1988

Structure and Genomic Organization of Two New Families of Human Repetitive DNA

Susan L. Carnahan
Loyola University Chicago

STRUCTURE AND GENOMIC ORGANIZATION OF TWO NEW FAMILIES OF HUMAN REPETITIVE DNA

by

Susan L. Carnahan

A Thesis Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of

Master of Science

March

1988

This thesis is dedicated to

three special friends.

ACKNOWLEDGEMENTS

I would like to thank my committee members, Howard Laten and John Smarrelli, for their time, effort and friendship. Special thanks is to be given to my advisor, Jeff Doering, for his patience. I would like to thank all my friends, old and new, who have seen me struggle and given me support.

VITA

The author, Susan Leah Carnahan, is the wife of James Richard Carnahan and mother of Faith Elizabeth and Elizabeth Ann. She was born on December 2, 1945 in Highland Park, Illinois.

Kindergarten and first grade education were received at a Catholic elementary school and the remainder of her elementary education was received in the public schools of Highland Park. Her secondary education was completed in June 1963 at Highland Park High School.

In June 1967 she was awarded a Bachelor of Science in Education with a biology major from Northern Illinois University. During the years following her graduation she taught in three public school systems, worked at a downstate hospital and clinic as a lab technician and the years following the birth of her second child she was a substitute teacher.

She was awarded an assistantship in biology in the fall of 1985 at Loyola University of Chicago, allowing her to complete the requirements for a Masters of Science in 1988.

In 1987 Susan Carnahan gave a presentation at the Graduate Research Forum sponsored by the Loyola University Chapter of Sigma Xi.

PUBLICATIONS

Doering, J.L., Burket, A.E. and Carnahan, S.L. (1988). A New Human Alphoid-like Repetitive DNA Sequence. FEBS Letters. (in press)

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

One of the themes of biology is the relationship of structure and function, but in the case of genomic DNA sequences the relationship can be difficult to ascertain. Many unique sequences of genomic DNA cause the production of proteins. However, one third of the human genome is made up of repetitive DNA (Britten and Kohne, 1968). While the detailed structure of many repeated sequences have been determined, at present no functions have been definitely assigned to them. Functions postulated for repetitive DNA include control of the transcriptional activity or inactivity of other DNA sequences (Wu and Manuelidis, 1980) or maintaining the basic structure of the chromosomes (Manuelidis, 1978b). Other theories suggest that repetitive DNA has no specific function at all (Orgel and Crick, 1980).

A family of repetitive DNA is made up of members with very similar repeat lengths and nucleotide sequences. These families can be grouped into two classes: interspersed repeats or tandem repeats (Lewin, 1980). Interspersed repeats can be found as single copies at various places throughout the genome (Singer, 1982), while tandem repeats are linked to each other in a chain like fashion (Lewin, 1980).

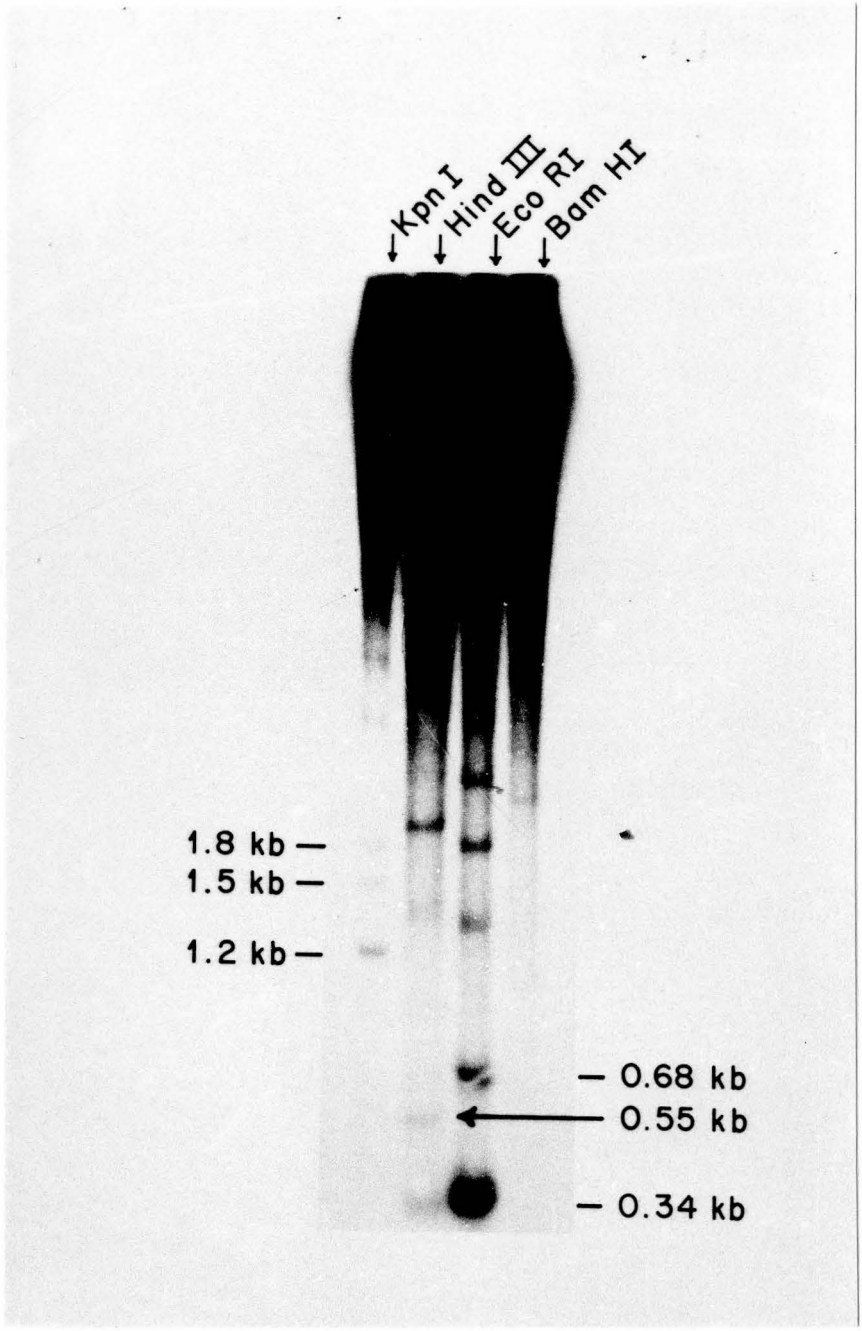
Much of the repetitive DNA in the human genome has still not been characterized, and new sequence families continue to be discovered. Previous work identified two

highly repetitive sequences 550 bp in length when total human DNA was digested with Hind III (Figure 1) (Doering and Burket, 1985). One sequence was an interspersed repeat and the other was a tandem repeat. The present work was undertaken in order to determine the detailed sequence organization of these two repeated sequences and to see if there was any relationship of the new families to previously described repetitive sequences.

The 550 bp interspersed sequence was found to be a variant of the larger interspersed sequence known as the Kpn I sequence (Hattori et al., 1985). This variant has 92.3% sequence similarity to a portion of the 5′ end of a full length Kpn I family member (Hattori et al., 1985). This is a much higher degree of sequence similarity than has been seen when other Kpn I family members are compared (Potter, 1984).

The new tandem repeat represents a previously undescribed family of alphoid sequences. Alphoid sequences have a basic repeating unit of 170 bp and are preferentially located at centromeres (Manuelidis, 1978b). They also have strong sequence similarity to the alpha component of African green monkey (AGM) DNA, from which their name is derived (Manuelidis and Wu, 1978). The new alphoid family has apparently been constructed from a repeating basic dimer of 340 bp. Members of this new family are unusually heterogeneous in sequence, and there are a num-

Figure 1. Genomic Digests Revealing Highly Repetitive Sequences.
Placental DNA was digested with the restriction enzymes indicated, blotted to nitrocellulose and probed with total genomic DNA under conditions where only repetitive sequences will be able to hybridize (see Materials and Methods). Sizes for selected fragments are indicated. The 0.55 kb Hind III fragments are denoted by the arrow.


Kpn I
Hind III
Eco RI
Bam HI
1.8 kb
1.5 kb
1.2 kb
0.68 kb
0.55 kb
0.34 kb

ber of variant sequence classes. Some of the variant classes exist in separate genomic domains and even on a single chromosome the members of such a class are not significantly mixed with members of another class. This new alphoid family also exhibits some chromosome specificity in its organization.

The present work provides clear evidence that some dispersed families of repetitive sequences are less heterogeneous in sequence than some tandem families. Thus previous concepts (Smith, 1976, Singer and Skowronski, 1985) concerning the maintenance of sequence homogeneity in repetitive DNA families may need to be revised.

REVIEW OF LITERATURE

There exists within the human genome a great variety of interspersed and tandemly repeated sequences (Singer 1982, Manuelidis and Wu 1978, Gray et al., 1985, Shimizu et al., 1983). The purpose of the present study is to compare two newly-described families of human repetitive DNA with those that have already been characterized.

Human Interspersed Sequence Families

Interspersed repeats are those sequence families of DNA whose individual members can be found as single copies at various places throughout the genome (Singer, 1982). There are two kinds of interspersed repeats classified according to size: SINEs (short interspersed repeated segments) and LINEs (long interspersed repeated segments) (Singer, 1982). SINEs are 500 bp or less and have a copy number of about 10^5 (Singer, 1982). The most well-characterized SINE in humans is the Alu family, which is named after the restriction enzyme that frequently cuts it. It has a copy number of 910,000 per haploid genome. This is a two-fold higher number than in other primates (Hwu et al., 1986). Approximately 3-6% of the human genome consists of the Alu family (Schmid and Jelinek, 1982). An individual repeat is 300 bp in length (Schmid and Jelinek, 1982). Each repeat is a dimer of two imperfect direct repeats with a 31-bp insertion in the second monomer (Deininger et al., 1981). The 3′ terminus

of the dimer is A-rich (Schmid and Jelinek, 1982). Alu repeats are randomly distributed throughout the genome with variable distances between them (Fritsch et al., 1980, Della Favira et al., 1981). Alu sequences may be situated close to other dispersed DNA sequences or to tandem repeats (Fritsch et al., 1980, Grimaldi and Singer, 1983). Another SINE is the O family whose members are 411 bp and 360 bp in length (Sun et al., 1984). This repeat is estimated to be 0.01% of the total human genome (Sun et al., 1984).

By definition a LINE is over 5 kilobases (kb) long (Singer, 1982). In humans the Kpn I family, named after the restriction enzyme that frequently cuts it, is the best-characterized LINE. The members of the Kpn I family, when found in full length, are approximately 6.4 kb (Shafit-Zagardo et al., 1982a). Such a copy is found at the 3′ end of the beta globin gene cluster in humans (Shafit-Zagardo et al., 1982a). Full-length members are made up of four subregions, defined by Kpn I fragments, whose linear order 5′ to 3′ is 1.8 kb, 1.5 kb, 1.2 kb, and 1.8 kb (Shafit-Zagardo et al., 1982a). Kpn I family, also known as L1Hs (LINE 1 Homo sapiens), members are found on all autosomes and the X chromosome (Shafit-Zagardo et al., 1982a), demonstrating its ubiquitousness. Restriction mapping studies demonstrate that L1Hs sequences contain no Alu sequences nor internal tandem repeats (Shafit-Zagardo

et al., 1982a). Thus L1Hs is a distinct repetitive sequence.

There is a great deal of heterogeneity in the structure of L1Hs family members. This complexity is seen from the following observations: 1) The linear order of Kpn I subregion fragments differs from one family member to another (Potter, 1984). 2) Deletions exist. For example, the 1.5 kb subregion is missing from Kpn A, a variant Kpn I family member (Potter, 1984). 3) Frequently the full length sequence is truncated. Copies may be missing regions from the 5′ and/or 3′ ends (Sun et al., 1984, Shafit-Zagardo et al., 1982a). Such truncated sequences have lengths of 2.7 kb or 3.4 kb which are made up of 1.2 kb and 1.5 kb subunits or 1.5 kb and 1.8 kb subunit respectively (Shafit-Zagardo et al., 1982b, Grosveld et al., 1981). 4) Kpn I family members that contain the 5′ 1.8 kb subregion may or may not have an internal 131 bp sequence (Hattori et al., 1985). The copy number for full-length L1Hs family members is 10^4 per haploid genome (Hattori et al., 1985), yet the copy number for the truncated sequences lacking the 5′ end is 107,000 (Hwu et al., 1986).

Another LINE has been isolated that is distinct from L1Hs and is refered to as L2Hs (Musich and Dykes, 1986). A 0.6 kb Kpn I fragment representing this family was purified and cloned (Musich and Dykes, 1986). When this clone was used to probe total genomic DNA, a smear was

observed with lengths ranging from 0.2 kb to greater than 24 kb (Musich and Dykes, 1986). This pattern indicates that the sequence is present at many loci in the genome. The restriction patterns observed by digestions with Kpn I, Alu I and Hae III indicate that this is a LINE not a SINE (Musich and Dykes, 1986). As with the L1Hs family there is no evidence that L2Hs contains any other types of repeats (Musich and Dykes, 1986). In the human genome L2Hs exhibits extensive population polymorphism, since no two individuals examined have identical patterns of L2Hs restriction fragments (Musich and Dykes, 1986).

Interspersed Repetitive Sequences and Transposition

The widespread distribution of LINEs and SINEs in the genome suggests that these sequences are mobile. In particular, dispersed sequences have many characteristics that suggest they are transposable elements (Flavell and Ish-Horowicz, 1981, Temin 1980). Transposition is the duplication and movement of one sequence of DNA from its original location in the genome to a new location. The first transposable elements were identified in maize by McClintock (McClintock, 1984). In maize, unstable mutations in kernel pigment genes were shown to be caused by mobile genetic elements (McClintock, 1984). Prokaryotic transposable elements often carry genes which confer antibiotic resistance (Calos and Miller, 1980). Ty in yeast (Boeke et al., 1985) and P elements in Drosophilia (Sprad-

ling and Rubin, 1982) are well-known eukaryotic transposons.

There are characteristics that are shared by all transposable elements (transposons), although quantitative variations may exist. Prokaryotic and eukaryotic transposable elements have perfect or nearly perfect inverted terminal repeats of approximately 20-40 bp (Calos and Miller, 1980). Another common characteristic is the duplication of a short sequence at the target site, one at each end of the inserted element.

Interspersed repeats share some characteristics with lower eukaryotic transposable elements, such as inverted terminal repeats and/or target site repeats (Flavell and Ish-Horowicz, 1981, Temin, 1980). Frequently, however, the inverted terminal repeats are missing. For example the Alu family members are flanked by short direct repeats, 7-20 bp, which have different sequences at each individual Alu locus (Schmid and Jelinek, 1982), but there are no inverted terminal repeats. The O family has an imperfect eleven nucleotide direct repeat flanking its 5′ end and its A-rich 3′ end (Sun et al., 1984). A truncated member of the L1Hs family, Kpn A, has been found inserted into a cluster of the tandemly repeating alphoid sequences (Potter, 1984). The insertion (transposition) is believed to have taken place after amplification of the tandem sequence because if it had occurred prior to that, the

inserted sequence would also have been amplified. If one could precisely excise Kpn A, the alphoid sequence member that had been interrupted would be identical to the other members of that cluster. Thus, the direct repeats are missing in this case. A 2.3 kb transposon-like element in human DNA contains flanking repeats and short direct target repeats (Paulson et al., 1985). The two flanking long terminal repeats (LTR) are members of the O family characterized by Sun et al. (1984) (Paulson et al., 1985).

A precise endonucleolytic mechanism is suggested to be involved in the genomic rearrangement of the subregions of LINE sequences (Soares et al., 1985). It has been found that the LINE sequence, related to L1Hs, in rat shares common end points with L1Hs, indicating that they may be under control of a specific enzyme (Soares et al., 1985). In order for an endonucleolytic enzyme to work it would be necessary for it to recognize one or more specific sets of sequences, but as yet no specific sequence signals have been identified (Soares et al., 1985).

One way in which transposition can be mediated is by retroposition, a process which uses RNA to make complementary DNA (cDNA). Viral retroposons carry their own copy of the reverse transcriptase gene which they use to transcribe a cDNA sequence (Varmus, 1982). After the cDNA is made it integrates into the genome. The yeast Ty element is known to move around its genome via an RNA

intermediate (Boeke et al., 1985). Alu and Kpn I sequences both have A-rich 3′ termini which are evidence for movement within the genome via a DNA copy of mRNA (Schmid and Jelinek, 1982, Singer and Skowronski, 1985, Lerman et al., 1983). It seems possible that the sequence from which all L1Hs family members were initially copied could have carried or still does carry the sequence which stimulates the active distribution (transposition) of the L1Hs sequence (Hattori et al., 1986). One of the open reading frames (ORF) of this family possesses significant sequence similarity to several viral RNA-dependent DNA polymerase genes (Hattori et al., 1986): Moloney murine leukaemia virus (Shinnick et al., 1981), human T-lymphotropic virus type I (Seiki et al., 1983) and Rous sarcoma virus (Schwartz et al., 1983). Thus the similarity between L1Hs sequences and reverse transcriptase genes gives some credence to the concept that interspersed sequences are capable of retroposition (Hattori et al., 1986). It is thought that the Kpn I family member located at the 3′ end of the beta globin gene family may have descended from an active L1Hs sequence that encoded for a reverse transcriptase, thereby causing its active transposition (Fujita et al., 1987). If this protein is produced by L1Hs, it could bind to its own mRNA and make cDNA copies that could be dispersed throughout the genome (Hattori et al., 1986). As of yet there is no direct evidence that L1Hs

sequences are undergoing transposition in contemporary populations.

Transcription of Interspersed Repeats

In eukaryotes RNA polymerase II is used to transcribe sequences that encode functional proteins. Alu sequences are transcribed by RNA polymerase III (Schmid and Jelinek, 1982), which is known to transcribe 5S RNA and tRNAs, and will therefore not code for a functional protein. As with other genes transcribed by polymerase III, Alu sequences have an internal promoter (Jagadeeswaran et al., 1981, Van Arsdell et al., 1981). Some members of the Alu family produce RNA transcripts of heterogeneous size. These are not the result of processing, but rather are the result of a unique initiation and multiple terminations mediated by RNA polymerase III on the DNA template (Hess et al., 1985). L1Hs contains signals for transcription and translation, such as the TATA box and start and stop codons (Hattori et al., 1985). Most of the transcription of the KpnI family is directed by RNA polymerase II (Shafit-Zagardo et al., 1983) and heterogeneous nuclear transcripts of 400 bp to 10 kb (Shafit-Zagardo et al., 1983, Sun et al., 1984, Lerman et al., 1983) are seen. The longest open reading frame consists of 268 amino acid residues (Hattori et al., 1985). It is thought that since both strands of L1Hs are transcribed, these are non-specific read-through transcripts of longer transcription units (Sun et al., 1984).

Apart from being transposable elements, interspersed repeats could have specific functions. It has been proposed that interspersed sequences could serve in a regulatory capacity (Britten and Davidson, 1969), such as to turn off a gene by inserting into an active gene's ORF. This is not to say that all members of a given family will perform the same function because the location of the individual member may render it nonfunctional (Orgel and Crick, 1980). The controlling elements of McClintock (1984), which interrupted the expression of a pigment gene, are such an example. Manuelidis (1982) speculates that interspersed sequences may play a role in defining discrete chromosomal domains apart from those located at or near the centromere. It is possible that some repetitive elements in combination with special proteins may provide recognition sites for chromosome replication and condensation (Manuelidis and Ward, 1982).

Evolution of Interspersed Sequences

It is thought that initially there was a small group of functional, conserved copies of the Kpn I sequence (Soares et al., 1985). If homogenizing mechanisms failed, then family members could start drifting in sequence (Soares et al., 1985). These mutations could then render the sequence nonfunctional. Therefore, the evolution of interspersed sequences, specifically the KpnI family, could be closely related to its lack of function. After

the loss of the ability to produce a given protein the mutations (divergence) in sequence could continue at random without selective pressure.

Members of the L1Hs family are more homogeneous within species than between species (Burton et al., 1986). However, there is no difference between monkey and human in the location of several Kpn I sites, suggesting a fixation before the divergence of the two primates from a common ancestor (Grimaldi and Singer, 1983). Truncation frequently occurs at the 5′ end of the L1Hs and it is here that there is a particularly high GC content (Hattori et. al., 1985). This difference in nucleotide content suggests that the 5′ end of the Kpn I sequence may have had an independent origin from other parts of the sequence (Hattori et al., 1985).

The two sequences, L1Hs and L2Hs, are thought to have had an independent origin and evolution, since the cross-hybridizations that were performed proved these to be two distinct LINEs (Musich and Dykes, 1986). The latter is detectable in gorilla DNA but not in that of chimps and lower primates (Musich and Dykes, 1986). Since it is present only in the higher primates, L2Hs is thought to have developed more recently in evolutionary history than L1Hs (Musich and Dykes, 1986).

While evolution of L1Hs has involved a process of divergence, there has also been some homogenization (Lee

and Singer, 1986). Lee and Singer (1986) reported that the LINE sequences on a single chromosome, CAE-19, from the African green monkey were characteristically different from those present in the genome as a whole. Thus, one or more mechanisms responsible for the homogenization of this dispersed family were more effective within that single chromosome than between that chromosome and the remainder of the genome. Homogenization appears to involve multiple mechanisms controlled by a variety of unknown factors (Lee and Singer, 1986).

Over time interspersed sequences were amplified, thus establishing their high copy number. The dispersal of repetitive sequences by insertion and deletion has apparently been taking place in the genome of higher primates for the last few million years (Hwu et al., 1986). Thus the existence of interspersed sequences and their movement could be a major potential source of genomic evolution (Hwu et al., 1986).

Tandemly Repeated Sequences

The other major class of repetitive sequences found in the human genome is tandemly reiterated DNA. Here the repeating units are linked "head to tail," thus forming clusters of the same or closely related sequences. All such sequences, regardless of how they are isolated, are presently referred to as "satellite" DNA (Prosser et al., 1986). The simple satellites are composed of repeats

whose lengths are 5-10 bp and are isolated by density gradient centrifugation (Corneo et al., 1970). Repeats of higher complexity have repeat lengths of 68-172 bp (Meneveri et al., 1985, Shimizu et al., 1983) and do not contain simpler units. They are isolated by the use of restriction endonucleases (Manuelidis, 1978a), and are named according to the enzyme that cuts the major unit of the family plus the length of the major unit.

One tandemly repeated sequence in humans is the Hinf family, which consists of repeat units of 319 bp (Shimizu et al., 1983). It is a dimer having two related but distinct subunits of 172 bp and 147 bp (Shimizu et al., 1983). There is also a Sau 3A repetitive family whose repeating unit is 68 bp (Meneveri et al., 1985).

Alphoid DNA is a tandemly repeated sequence that is the most abundant and well-characterized in the human genome. The sequence is referred to as alphoid because of its sequence similarity to the alpha component of African green monkey (AGM) (Manuelidis and Wu, 1978) which is a tandem repeat of 170 bp. The alpha sequence in AGM comprises 13-20% of the genome (Maio, 1971, Fittler, 1977, Singer, 1979) and is found in the centromeric heterochromatin (McCutchan et al., 1982). The general definition of an alphoid sequence is that the sequence have a 170 bp repeat and be located in the centromeric region.

Alphoid Sequence Families

There are a number of alphoid sequence families within the human genome. The organization of a given family is characterized by 1) length of the amplified repeat unit, 2) particular restriction enzyme(s) used to visualize the repeat in digests of genomic DNA and 3) primary DNA sequence (Willard et al., 1986).

The first human alphoid sequences studied were detected by digesting total genomic DNA with Eco RI and staining with ethidium bromide (Manuelidis, 1978a). This revealed bands that were 340 bp and 680 bp in length (Manuelidis, 1978a), integral multiples of 170 bp. The Eco RI 340 bp sequence has become the standard to which all other alphoid sequences are compared. This unit is a dimer consisting of two subunits of 169 bp and 171 bp, which differ in sequence by 27% (Wu and Manuelidis, 1980). There is a junction between the two subunits that due to mutation no longer contains an Eco RI restriction site (Wu and Manuelidis, 1980). The 680 bp repeat consists of two 340 bp units which show a divergence of less than 1% (Wu and Manuelidis, 1980). It is estimated that the Eco RI family (340 and 680 bp sequences) make up 1% of total human DNA and is present in over 100,000 copies per cell (Furlong et al., 1986).

Using the technique of *in situ* chromosomal hybridization, the Eco RI 340 bp sequences were shown to be at the

centromeric region of all chromosomes, with the highest concentrations on chromosomes 1, 3, 7, 10, and 19 (Manuelidis, 1978b). It should be noted that complex repeating units and simple satellites can occupy the same general region of a given chromosome (Manuelidis, 1978b). While it is known that the alphoid sequences, in particular the Eco RI 340, are found in the centromeric heterochromatin of all chromosomes (Manuelidis, 1978b), there is no indication that the centromeric heterochromatin of all human chromosomes also contains simple satellite DNA sequences (Gosden et al., 1975).

Within the Eco RI 340 bp family there exists a significant amount of heterogeneity. One study, using 24 clones that contained alphoid Eco RI 340 bp fragment inserts, showed that no two sequences were exactly alike, and the average divergence from exact homology was 5.2% (Furlong et al., 1986). A similar study of 45 cloned Eco RI 340 bp fragments revealed the existence of at least 20 distinct sequence subfamilies, some of which share regions of conserved sequences (Jorgensen et al., 1986). This study suggested that the number of Eco RI 340 bp subfamilies may be equal to or exceed the number of chromosomes (Jorgensen et al., 1986).

In addition to the Eco RI 340 bp family there are many other alphoid sequences, some of which can be found on all chromosomes. These other alphoid families are all clearly

related in sequence to the Eco RI 340 bp family and are often organized as higher-order repeats. These repeats are frequently made up of divergent monomers, which appear to have been amplified from a prototype monomer. These monomers remain linked to each other and then the collective unit is tandemly repeated. A Sau 3A alphoid family, which is found at numerous places in the human genome, including extrachromosomal DNA, has an 849 bp repeat consisting of five relatively homologous tandem subunits of 171 bp, 171 bp, 167 bp, 169 bp and 171 bp (Kiyama et. al., 1986). Comparison of these subunits with subunits I and II of Eco RI 340 bp (Wu and Manuelidis, 1980) yielded an average sequence similarity of 70.5% and 73.7%, respectively. Gillespie et al. (1982) have identified a 340 bp Xba I alphoid tandem repeat which contains two 170 bp subunits substantially different in sequence from the Eco RI 340 bp family. The 308 family (Jabs et al., 1986) is a 3.0 kb Bam HI human DNA fragment that is present in the centromeric region of all human chromosomes and cross-hybridizes to other alphoid sequences. Population polymorphism within this sequence can be observed with several restriction enzymes (Jabs et al., 1986). Gray et al. (1985) have sequenced yet another alphoid family, the Xba 682 bp sequence, that can be found at the centromeric region of all human chromosomes except the Y chromosome. The average mismatch between the 170 bp subunits of this

sequence is greater than 32%, but between alternating subunits the mismatch is only 21% (Gray et al., 1985). Between the Eco RI 680 bp and the Xba I 682 bp sequences the mismatch is 21.2% (Gray et al., 1985). Another alphoid sequence, p82H, that can be hybridized to centromeres of all the human chromosomes is composed of 14 tandemly repeated variants of a basic 172 bp sequence (Mitchell et al., 1985). Within the 14 units there is a 8-26% sequence mismatch (Mitchell et al., 1985). These units have 37-51% mismatch with subunit I of the Eco RI 169 bp (Wu and Manuelidis, 1980) and 45-59% mismatch with the second subunit (Mitchell et al., 1985). The grains seen during the in situ hybridization of p82H are restricted to the centromere region and more evenly distributed among the chromosomes than Xba I family (Gray et al., 1985) and Eco RI family (Wu and Manuelidis, 1980).

Alphoid DNA on different chromosomes usually cross-hybridizes to some degree (Manuelidis, 1978b). The degree of hybridization stringency allows for the detection of greater or lesser mismatch of DNA sequences. A sequence that appears chomosome-specific at higher stingency may lack that specificity at reduced stringency. For example, at reduced stringency, a 2.0 kb Bam HI alphoid fragment binds to the centromeres of all autosomes and sex chromosomes (Willard, 1985) while at a higher stringency it appears to be specific to the X chromosome. Thus, in

addition to the alphoid families found on all chromosomes in the genome there are others which are chromosome specific. The 2.0 kb sequence specific to the X chromosome is comprised of 12 divergent monomers, each of which is related in sequence to the prototype primate alpha satellite sequence (Waye and Willard, 1985). On the X chromosome there are 5000 copies of the 12-monomer repeat unit (Willard et al., 1986). The individual monomer units have 65-85% sequence similarity to each other (Willard et al., 1986), while the independent higher-order units have greater than 99% sequence similarity (Waye and Willard, 1986). A 16-monomer higher-order repeat unit, approximately 2.7 kb, has 1000 copies on chromosome 17 (Willard et. al., 1986). Mismatches of 15-40% exist between monomers of the 16-mer, but are less than 2% between independent 16-mer higher order repeats (Willard et al., 1986). The 2.7 kb family on chromosome 17 exhibits a good deal of population polymorphismfor restriction sites within the repeat units (Willard et al., 1986). There is an alphoid repeating unit of 5.7 kb specific for the Y chromosome (Tyler-Smith and Brown, 1987). The size of the alphoid blocks containing this repeat differ by 100 kb in the two individuals examined, but the structure of the flanking DNA appears to be similar (Tyler-Smith and Brown, 1987). Within the 5.7 kb repeating units are 170 bp subunits, which have a sequence similarity between 76% and 86%

among themselves (Tyler-Smith and Brown, 1987).

Divergent alphoid satellite monomer repeats can be found on small polydispersed circular DNA which is ubiquitous in eukaryotic cells (Jones and Potter, 1985). This extrachromosomal alphoid DNA is thought to be derived from alphoid sequences on the chromosomes (Kiyama et al., 1986) and might act as an intermediate for moving alphoid sequences to new locations. Such a mechanism would explain how identical satellite sequences can exist on non-homologous chromosomes and at non-centromeric locations (Jones and Potter, 1985).

Function of Alphoid Families

All alphoid sequences examined to date are located in the centromeric region (Wu and Manuelidis, 1980, Gray et al., 1985, Waye and Willard, 1985 and Jorgensen et al., 1986). Fitzgerald-Hayes et al. (1982) saw a sequence similarity of 72% between the Eco RI alphoid sequence and the known functional centromere Element I region of yeast DNA, suggesting a possible centromeric function for the alphoid family. Other hypothesized functions for alphoid DNA include the creation of higher order chromosome structure (Furlong et al., 1986) and the phasing of nucleosome positions (Maio et al., 1977). In fact Strauss and Varshavsky (1984) have found a protein that binds to the alpha component of AGM DNA and which could be responsible for the nucleosome phasing on that particular sequence. It

is also possible that alphoid repeats could serve a regulatory function (Davidson and Britten 1973). Wu and Manuelidis (1980) have suggested that these reiterated sequences might determine the transcriptional activity or inactivity of a region. Furlong et al. (1986) propose that alphoid sequences could regulate recombinational events.

The original Eco RI 340 bp consensus sequence contains stop codons in all reading frames (Manuelidis and Wu, 1978), thus making a protein coding function unlikely. Orgel and Crick (1980) suggest that such repetitive DNA is "parasitic," existing without benefit to the genome, and with no feasible function. Jorgensen et al. (1986) suggest that if some distinct features of the repeat units were conserved, these features might give some insight into a possible function for the repeat.

Evolution of Alphoid Sequences

There are two universal features of the alphoid family that have been preserved through the evolution of this family: the length of the basic repeat, 170 bp, and its tandem arrangement (Wu and Manuelidis, 1980). The lack of homology between highly repeated simple satellite DNA and the Eco RI fragments (Manuelidis, 1978a) is a strong indication that although these two families are both highly repetitive and are tandemly repeated they shared no sequence similarity at any time during the

evolutionary process (Wu and Manuelidis, 1980).

The evolution of tandemly repeated sequences is thought to include the process of homogenization, which includes unequal crossing over (Smith, 1976). This process involves the out-of-register alignment of tandem repeats on the paired chromosome homologs, which can occur because only short segments of similar sequence are required for pairing. After the unbalanced alignment, crossing-over will result in tandem clusters of unequal length. After repeated rounds of such out-of-register crossing-over sequences with high sequence similarity tend to remain adjacent to each other in the middle while at the ends of the tandem array the sequences are more dissimilar (Smith, 1976).

The establishment of the Eco RI 340 bp family is proposed to have occurred in two steps; dimer formation from divergent monomers and amplification of dimer into long tandem arrays (Manuelidis and Wu, 1980). A similar multistep process is thought to have taken place in the 2.0 kb fragment, a 12-mer, on the X chromosome, beginning with the formation of a six-monomer unit and then unequal crossing-over of two six-monomer units (Waye and Willard, 1985).

There are a large number of alphoid families that evolved from a prototype monomer. During evolution sequences were amplified laterally and then unequal crossing-

over could take place. These changes altered the length of the higher order repeats. Throughout the entire process sequences could be mutated to alter restriction sites creating new sites and deleting previously existing ones. When divergence is observed amplification is thought to have taken place after the mutation of a restriction site, since if one looks at a cluster of repeats the mutation is seen in the analogous position of all monomers (Lee and Singer, 1982). After divergence some regions will remain conserved acting as an indicator of their ancestry. Within the 12 repeating units of the 2.0 kb alphoid sequence on the X chromosome several domains, each 7-25 bp in length, have been conserved (Waye and Willard, 1985). In a comparison study (Waye and Willard, 1985) the same regions remain conserved within the alphoid Eco RI 170 bp repeating unit of human (Wu and Manuelidis, 1980), baboon (Donehower and Gillespie, 1979) and African green monkey (Rosenberg et al., 1978).

The various alphoid families are not present in all primates. It appears that the Eco RI family was the first alphoid sequence amplified after humans, apes and gibbons diverged from Old World monkeys (Lai et al., 1979). After humans and apes diverged from gibbons the 340 bp Xba I satellite was established (Gillespie et al., 1982). In the time span since humans diverged from great apes the 682 bp Xba I satellite has amplified (Gray et al., 1985).

The distribution of man's tandemly repetitive DNA is unusual compared to that of other mammals (Mitchell et al., 1985). Other mammals have their major tandemly repeated sequences equally distributed among their chromosomes in the pericentric heterochromatin (Pardue and Gall, 1970, Kurnit and Maio, 1973). Man's alphoid DNA shows chromosome specificity and thus is unequal in its distribution.

MATERIALS AND METHODS

Sources of DNA

Placental DNA, purified as in Blin and Stafford (1976) was a gift of D. Bieber. DNA from hamster-human hybrid cell line 153-E9A (Moore et al., 1977), containing human chromosome 21, was a gift from M. Cummings.

The following plasmids containing sequences of human repetitive DNA were used in this study: pHH550-2, a dispersed repeat, pHH550-31, a tandemly repeating sequence (Doering and Burket, 1985), and pHE340-64, a standard 340 bp alphoid repeating sequence (Doering et al., 1986).

Restrictions Digests

Restriction enzymes were purchased from Bethesda Research Laboratories (BRL), and the manufacturer's recommended buffers were used for the digest reactions, which were performed at 37° C for 2.5 hrs giving complete digestion. Restriction digests, containing 0.5 ug of 153-E9A DNA or 1 ug of placental DNA, were run on gels and blotted. Large scale digests, containing 25-75 ug of plasmid DNA, were used for the sequencing reactions.

DNA Transfer

Fractionation of restricted DNA was done by electrophoresis on 1% agarose gels by standard techniques (Doering et al., 1982, Sims et al., 1983). The DNA was transferred to Gene Screen Plus (NEN Research Products) using the alka-

line method (Reed and Mann, 1985). All probes were labelled with ^{32}P by nick-translation (Rigby et al., 1977). The membranes were pre-hybridized for 5 hours at 37° C. in hybridization solution (50% formamide, 1 M NaCl, 50 mM Tris [pH 7.5], 1% sodium dodecyl sulfate (SDS), 10 ug/ml denatured E. coli DNA). The denatured probe, containing 0.25-0.5 ug of DNA, was added and hybridization was allowed to proceed overnight at 37°C. Unless otherwise noted, the membranes were washed twice at room temperature for 10 minutes with 2X SSC (1X SSC is 150 mM NaCl, 15 mM sodium citrate, 0.1 mM EDTA), twice at 60° C. for 30 minutes with 2X SSC plus 1% SDS and twice at room temperature for 30 minutes with 0.5X SSC. The membranes were then air-dried and autoradiographed using Kodak XAR film.

Restrictions Prior to End-labelling

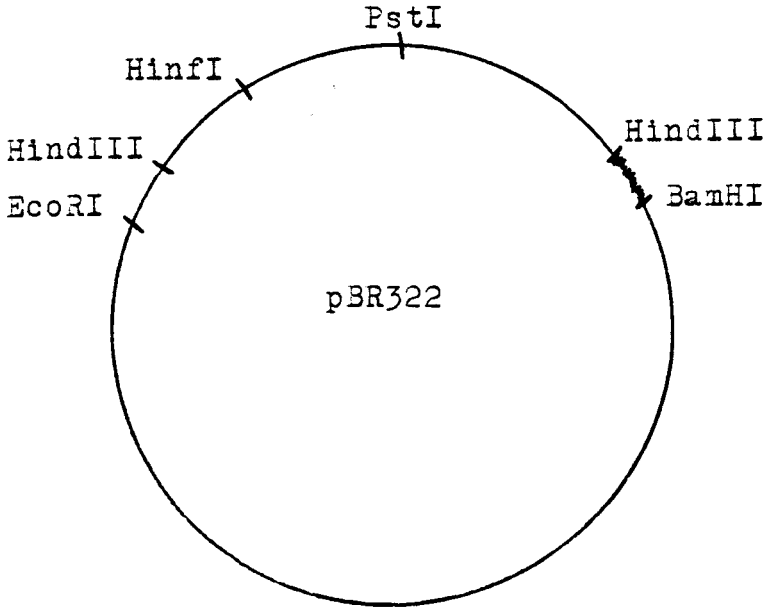
To sequence pHH550-2 it was necessary to do separate end-labellings in which the plasmid was initially cut with either Eco RI, Pst I or Hind III (see Figure 2 and Results for detailed sequencing strategy). In order to sequence pHH550-31 the plasmid was initially digested with either Eco RI, Bgl II or Hind III.

End-labelling

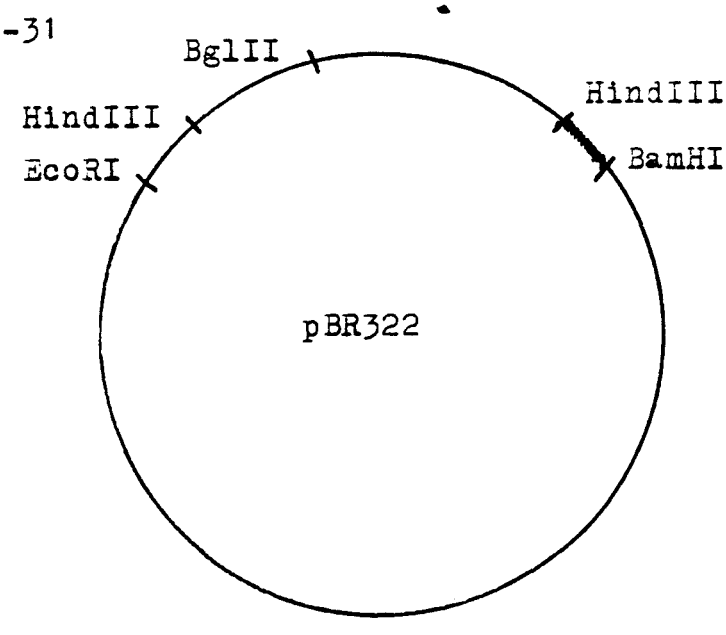
Prior to the bacterial alkaline phosphatase (BAP) treatment, digested plasmids were purified over NACS PREPAC columns (BRL) according to the manufacturer's directions. DNA amounts of 8.6-65 ug were redissolved in

Figure 2. Maps of Plasmids pHH550-2 and pHH550-31. Each plasmid (pBR322) contains an insert of 550 bp at the Hind III site. The drawing is not to scale and the wavy line represents a long length of DNA. See text (Methods and Materials and Results) for sequencing strategy. Selected restriction sites are shown.

a. pHH550-2



b. pHH550-31



17.5-35 ul of 0.2X SET (1X SET is 150 mM NaCl, 50 mM Tris pH 7.9, 1 mM (ethylenedinitrilo)-tetraacetic acid disodium salt [EDTA]) and then incubated at 65° C. for 60 minutes with 200-370 units of BAP (BRL) in order to remove the 5′ phosphate groups. The reaction was stopped with a 10% volume of 10 mM phenanthroline and then incubated at 65° C. for 30 minutes. The reaction was diluted to 100 ul with 0.2X SET and brought to a final concentration of 1% SDS, after which two phenol extractions were done. Sodium acetate was added to give a final concentration of 0.3 M and then three volumes of 95% EtOH were added. DNA was precipitated overnight at -80° C. The suspension was then centrifuged for 15 minutes at 12,000 rpm in a Microfuge (Beckman), at 4° C. The pellet was washed twice with 70% EtOH, vacuum-dried and redissolved in 14.5-18.5 ul of 0.2X SET. This DNA was used in a kinase reaction which contained 0.33-0.5 mCi of gamma-labelled crude ^{32}P ATP (ICN) and 20-30 units of T4 kinase (BRL) in 1X kinase buffer (1X kinase buffer is 70 mM Tris pH 7.5, 10mM $MgCl_2$, 0.1M KCl and 0.5mM DTT). The reaction was run at 37° C. for 30 minutes and stopped by being brought to a final concentration of 0.5% SDS and 10 mM EDTA. After the volume was raised to 50 ul with 0.2X SET, phenol extraction was done. Labelled DNA was separated from unincorporated ATP by P60 column chromatography in 0.2X SET. Butanol extraction (Maniatis et al., 1982) was done to reduce the collected

sample volume to 100-200 ul.

Gel Elution

The 550 bp fragments are in the vector pBR322, from which they must separated (Figure 2). Before the actual sequencing reactions could be done it was necessary to separate labelled ends by restriction digests. Following end-labelling pHH550-2/Eco RI (cut with Eco RI prior to end-labelling) was cut with Bam HI, pHH550-2/Pst I was cut with Bam HI and Eco RI, and pHH550-2/Hind III was cut with Hinf I. The procedure for pHH550-31 was the same as for pHH550-2 except the end-labelled fragments were cut as follows (Figure 2 and Results): pHH550-31/Eco RI was cut with Bam HI, pHH550-31/Bgl II was cut with Bam HI and pHH550-31/Hind III was cut with Bgl II. Unless otherwise noted, the digests were run out on 1% agarose gels. The gel was wrapped in plastic and exposed to Kodak XAR film. The film was used as a template for the excision of the desired fragment. Elution took place overnight in 6-8 ml of gel elution buffer (0.5M NH_4 acetate, 0.01M Mg acetate, 0.1% SDS, 0.1mM EDTA) at 37° C. The eluate was filtered through glass wool and the remaining gel was eluted again for 2 hours with an additional 2 ml of gel elution buffer. This second eluate was passed through glass wool, pooled with the first eluate and passed through a 0.22 um Millipore filter unit (Millex-GS). Three volumes of 100% EtOH were added and precipitation took place at -20° C for

2-3 days. DNA was pelleted by ultracentrifugation at 30 k for 30 minutes at 4° C in an SW41 rotor (Beckman). The pellet was washed twice with 70% EtOH and vacuum dried. It was redissolved in a pooled total of 600 ul of 0.3 M Na acetate, three volumes of 95% EtOH were added and then the suspension was placed at -80° C overnight. The precipitate was collected by centrifugation at 12,000 rpm in the Microfuge for 15-30 minutes at 4° C. The pellet was redissolved in 25 ul of H_2O. It was then reprecipitated with 95% EtOH for a minumum of 2 hrs. at -80° C. DNA was then again collected by centrifugation. The pellet was again washed twice with 70% EtOH, vacuum dried and redissolved in 33-65 ul H_2O.

Sequencing

Sequencing proceeded according to Maxâm and Gilbert (1980). The overall technique involves nucleotide specific reactions that modify the base, displace the modified base and then break the DNA. In the reaction specific for thymines and cytosines, hydrazine attacks at the C4 and C6 positions, opens the pyrimidine ring, and cyclizes with C4-C5-C6 to form new hydrazine-containing five-membered rings. Extensive hydrazine reactions remove these modified bases from the sugars. Piperidine, at 90° C, breaks DNA only at sugars without bases by eliminating both phosphates from the sugar. In the presence of high salt the reaction of hydrazine with thymine is inhibited, thus

giving a cytosine-specific cleavage reaction. Dimethylsulfate (DMS) methylates the N7 of guanine in DNA, fixing a positive charge into the N7-C8-N9 imidazole portion of the purine ring. Piperidine, as the free base, then opens the ring, displaces it from the sugar backbone and ultimately breaks the DNA at the resulting empty sugar. Acid depurination modifies and displaces guanine and adenine, but makes no distinction between the two. The purine rings are protonated and then eliminated from the sugar backbone by acid treatment. Piperidine will again break the DNA at the empty sugars. The four reactions thus cleave DNA at either C, C + T, G or A + G. By running the products of these reactions in adjacent gel lanes the nucleotide sequence can be readily determined. In the reactions that modify both prymidines or both purines it is necessary to double the quantity of DNA used. The procedures were performed as directed by Maxam and Gilbert (1980) except that the reaction times were shortened as follows to permit reading farther into the sequence; guanine: 3.5 minutes, cytosine: 4.0 minutes, and cytosine plus thymine: 4.0 minutes.

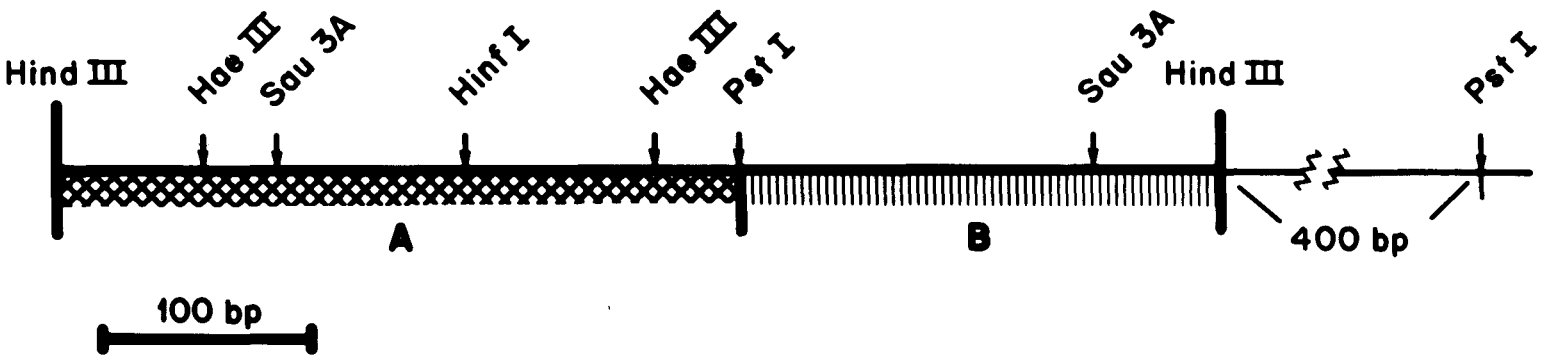
RESULTS

Previous work employed a technique that only reveals sequences of DNA that are highly repetitive (Figure 1) (Doering and Burket, 1985). Placental DNA was digested with a variety of restriction endonucleases, blotted and probed with total genomic DNA. Repetitive Hind III 550 bp fragments were detected that have not previously been described. When these fragments were cloned, cross hybridization of the different clones revealed the presence of two distinct sequences each 550 bp long. The sequences were designated pHH550-2, an interspersed sequence, and pHH550-31, a tandem repeat.

pHH550-2

Genomic DNA cut with a variety of enzymes was probed with pHH550-2. The blots showed smears with superimposed bands that were not integral multiples of any given monomer. These results indicated that pHH550-2 is an interspersed sequence that is located on fragments of varying length, especially higher molecular weight DNA (Doering and Burket, 1985). The restriction map (Figure 3) showed some similarities to a fully-characterized member of the major dispersed family, L1Hs. This member (T-beta-G41) is found downstream of the 3′ end of the beta globin gene cluster (Shafit-Zagardo et al., 1982a). In particular, the spacing between the Pst I site and the Hind III site appeared to be the same (Hattori et al., 1985) in pHH550-2

Figure 3. Restriction Map of pHH550-2.
Locations of restriction sites for several enzymes are indicated along with the size scale. Restriction enzymes that do not cut in this sequence are Alu I, Hha I, Hpa II and Taq I.


Hind III
Hae III
Sau 3A
Hinf I
Hae III
Pst I
Sau 3A
Hind III
Pst I
A
B
400 bp
100 bp

and T-beta-G41. Prior work also consisted of genomic digests using Kpn I or Pst I that were then probed with pHH550-2. This detected bands 1.9 kb for the former and 0.6 kb and 0.9 kb for the latter (Doering and Burket, unpublished data). These bands are similar in length to those found as part of T-beta-G41.

Sequencing of pHH550-2 was done to see if it was indeed a variant member of the L1Hs family. A representative sequencing gel is shown in Figure 4. All end-labelling was done at the 5′ end. Therefore, all sequences were read 5′ to 3′. The following sequencing strategy (Figure 5) was used: 1) The Eco RI site, outside the insert, was labelled and 175 bp were read from the left end of the insert in the direction of the Pst I site. 2) Next the Pst I site was labelled. Reading the opposite strand, 190 bp was determined from the Pst I site toward the Eco RI site. This created an overlap of 25 bp with the sequence read from the Eco RI site. 3) Finally, the right hand Hind III site was labelled. Reading from this site back to the left a total of 310 bp of sequence were determined. This overlapped the Pst I site by 90 bp and thus completed the full sequence.

The sequence of pHH550-2 (Figure 6) was found to be 547 bp long and is A + T rich (59.2%) with no internal repeats. It was then necessary to determine if pHH550-2 was a variant member of the L1Hs family. This verifica-

Figure 4. Representative Sequencing Gel.
This is the higher molecular weight portion of a short run (4-6 hrs) of an 8% polyacrylamide sequencing gel of pHH550-31 end-labelled at the Eco RI site. The lanes, left to right, are A + G, G, C and C + T. Positions 23-94 (Figure 16) can be clearly read up the ladder.

Figure 5. Sequencing Strategy for pHH550-2.
Arrows indicate direction and length of sequencing from a particular restriction site. See text for detailed sequencing strategy.

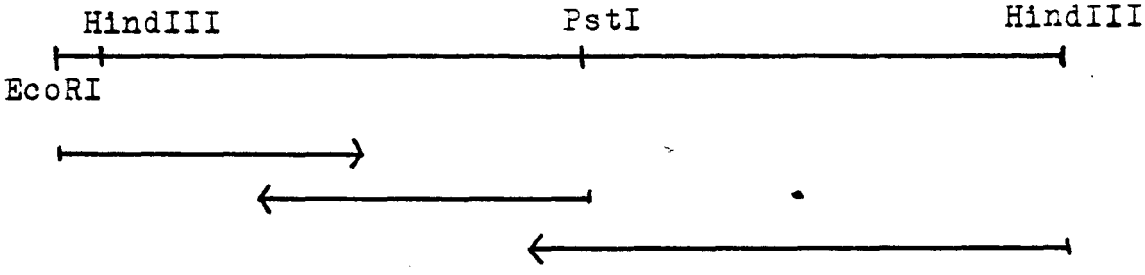

HindIII
PstI
HindIII
EcoRI

Figure 6. Sequence of pHH550-2.
The nucleotide sequence of pHH550-2 with the restriction sites corresponding to those on the restriction map (Figure 3).

```
        *         *         *         *         *
       10        20        30        40        50
GCTTCAGTAGCCAATTCGATCAACATGGAAGAAAGGGTATCAGTGATGGA

        *         *         *         *         *
       60        70        80        90       100
AGATCAAATGAATGAAATGAAGCGAGAACAGAAGTTTAGAGAAAAAAGAG
 Sau3A

        *         *         *         *         *
      110       120       130       140       150
TAAAAAGAAATAAACAAAGCCTCCAAGAGATATGGGACTATGTGAAAAGA

        *         *         *         *         *
      160       170       180       190       200
CCAAATCTATGTCTGATTGGTGTACCTGAAAGTGATGGGGAGAATAGAAC

        *         *         *         *         *
      210       220       230       240       250
CAAGTTGGAAAACACTCTGCAGGATATTATCCAGGAGAACTTCCCCAACC
                PstI

        *         *         *         *         *
      260       270       280       290       300
TAGCAAAGCAGGCCAACATTCAAATTCAGGAAATACAGAGAATGCCACAA
          HaeIII

        *         *         *         *         *
      310       320       330       340       350
AGATACTCCTTGAGAAGAGCAACTCCAAGACACATAATTGTCAGATTCAC
                                          HinfI

        *         *         *         *         *
      360       370       380       390       400
CAAAGTTGAAATGAAGGAAAAAATGTTAAGGGCAGCCAGAGAGAAAGGTC

        *         *         *         *         *
      410       420       430       440       450
GGGTTACCCACAAAAGGAAGCCCATCAGACTAACAGCGGATCTCTTGGCA
                                    Sau3A

        *         *         *         *         *
      460       470       480       490       500
GAAACTCACAAGCCAGAAGAGAGTGGGGGCCAATATTCAACATTCTTAAA
                          HaeIII

        *         *         *         *         *
      510       520       530       540       550
GAAAAGAATTATCAACCCAGAATATCATATCCAGCCAAATTAAGCTT
                                          HindIII
```

tion was done through a sequence comparison (Figure 7) of pHH550-2 with T-beta-G41 (Hattori et al., 1985) and with another variant L1Hs sequence, Kpn A (Potter, 1984), which has an unusual permutation and a deletion. The orientation of the sequence of pHH550-2 in Figure 7 has been reversed with respect to that in Figure 6 in order to facilitate its alignment with T-beta-G41 (Hattori et al., 1985). pHH550-2 was found to be part of the 5′ 1.8 kb subregion of T-beta-G41 (Figure 8). The comparison revealed a sequence similarity of 92.3% between pHH550-2 and the analogous region of T-beta-G41. The comparison with Kpn A shows an overall sequence similarity of 48.4%. Near the 5′ end, as it is oriented in Figure 8, positions 26-219 of pHH550-2 and Kpn A have a sequence similarity of 92.6%. A rearrangement of Kpn A has taken place 3′ of position 220 (Potter, 1984) causing this 3′ region to contain sequences from other portions of L1Hs that are not analogous to those of pHH550-2. There are six open reading frames in L1Hs, but pHH550-2 is located 5′ to all of them (Hattori et al., 1985).

pHH550-31

The second cloned fragment to be characterized, pHH550-31, represents a tandemly repeated sequence. Genomic DNA was digested with restriction enzymes that do not cut within the pHH550-31 sequence and probed with pHH550-31 (Figure 9). The five left lanes (Figure 9) show that

Figure 7. Comparison of the Sequences of T-beta-G41, pHH550-2 and Kpn A.
The orientation here has been reversed with respect to Figure 6 in order to facilitate alignment with T-beta-G41 (Hattori et al., 1985). The analogous sequence of Kpn A (Potter, 1984) is aligned with T-beta-G41. pHH550-2 is found in the 5′ 1.8 kb subregion of T-beta-G41.

```
                *         *         *         *         *
               10        20        30        40        50
KpnI    C       A  G  GA        X                   A   C
-2      GCTTCAGTAGCCAATTCGATCAACATGGAAGAAAGGGTATCAGTGATGGA
KpnA    AGCCTCAGGAGATGA GTGATC AC           AA

                *         *         *         *         *
               60        70        80        90       100
KpnI       G                    AT   G                   A
-2      AGATCAAATGAATGAAATGAAGCGAGAACAGAAGTTTAGAGAAAAAAGAG
KpnA       G                          GG                T X

                *         *         *         *         *
              110       120       130       140       150
KpnI             CG                  A
-2      TAAAAAGAAATAAACAAAGCCTCCAAGAGATATGGGACTATGTGAAAAGA
KpnA               G G               A

                *         *         *         *         *
              160       170       180       190       200
KpnI             CA   A         G                     G
-2      CCAAATCTATGTCTGATTGGTGTACCTGAAAGTGATGGGGAGAATAGAAC
KpnA             C                              C         G

                *         *         *         *         *
              210       220       230       240       250
KpnI                                                     T
-2      CAAGTTGGAAAACACTCTGCAGGATATTATCCAGGAGAACTTCCCCAACC
KpnA                 T      T TA  CCAGAAT TACA TGAA T A

                *         *         *         *         *
              260       270       280       290       300
KpnI         G         X     C                   C
-2      TAGCAAAGCAGGCCAACATTCAAATTCAGGAAATACAGAGAATGCCACAA
KpnA    A TTA CAAGAAAA  ACAAAC  CC CATCG A ACTG GCAAAGGAC

                *         *         *         *         *
              310       320       330       340       350
KpnI             A     A                     C  A
-2      AGATACTCCTTGAGAAGAGCAACTCCAAGACACATAATTGTCAGATTCAC
KpnA    T  ACAGA ACTTCTCA AAG AGA  TTTATGCAGCCAAAA ATACATG

                *         *         *         *         *
              360       370       380       390       400
KpnI
-2      CAAAGTTGAAATGAAGGAAAAAATGTTAAGGGCAGCCAGAGAGAAAGGTC
KpnA    A   AA TCTCACC TC CTGGCCA C GA AA TG  A TCA  CCA
```

```
                *         *         *         *         *
              410       420       430       440       450
KpnI                  G                     T     A C
-2     GGGTTACCCACAAAAGGAAGCCCATCAGACTAACAGCGGATCTCTTGGCA
KpnA   AATGAGATAC  TCTCAC C AGT AGA TGGTG T ATTAAAAAGTCGG

                *         *         *         *         *
              460       470       480       490       500
KpnI          T             A                       G
-2     GAAACTCXACAAGCCAGAAGAGAGTGGGGGCCAATATTCAACATTCTTAA
KpnA        AAC GGTG TG  GAG  TGT  AGAA TAGGAAC  T  TAC C

                *         *         *         *         *
              510       520       530       540       550
KpnI             T  GG        T               C
-2     AGAAAAGAATTATCAACCCAGAATATCATATCCAGCCAAATTAAGCTT
KpnA   T TTGGTGGGAC GT AA TAG  CAACC  TGT GA GTCAGT TGG
```

Figure 8. Positioning of pHH550-2 in T-beta-G41. pHH550-2 is found as part of the 5′ 1.8 kb subregion of T-beta-G41, the L1Hs full-length family member that is found downstream of the beta globin gene cluster (Shafit-Zagardo et al., 1982a). The filled in box shows the position of pHH550-2.

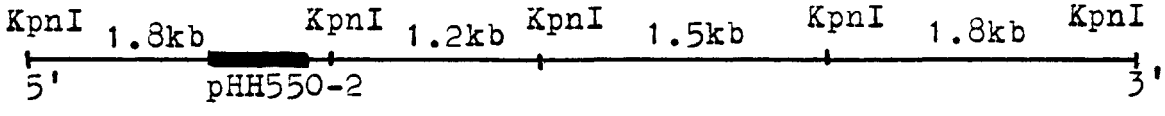
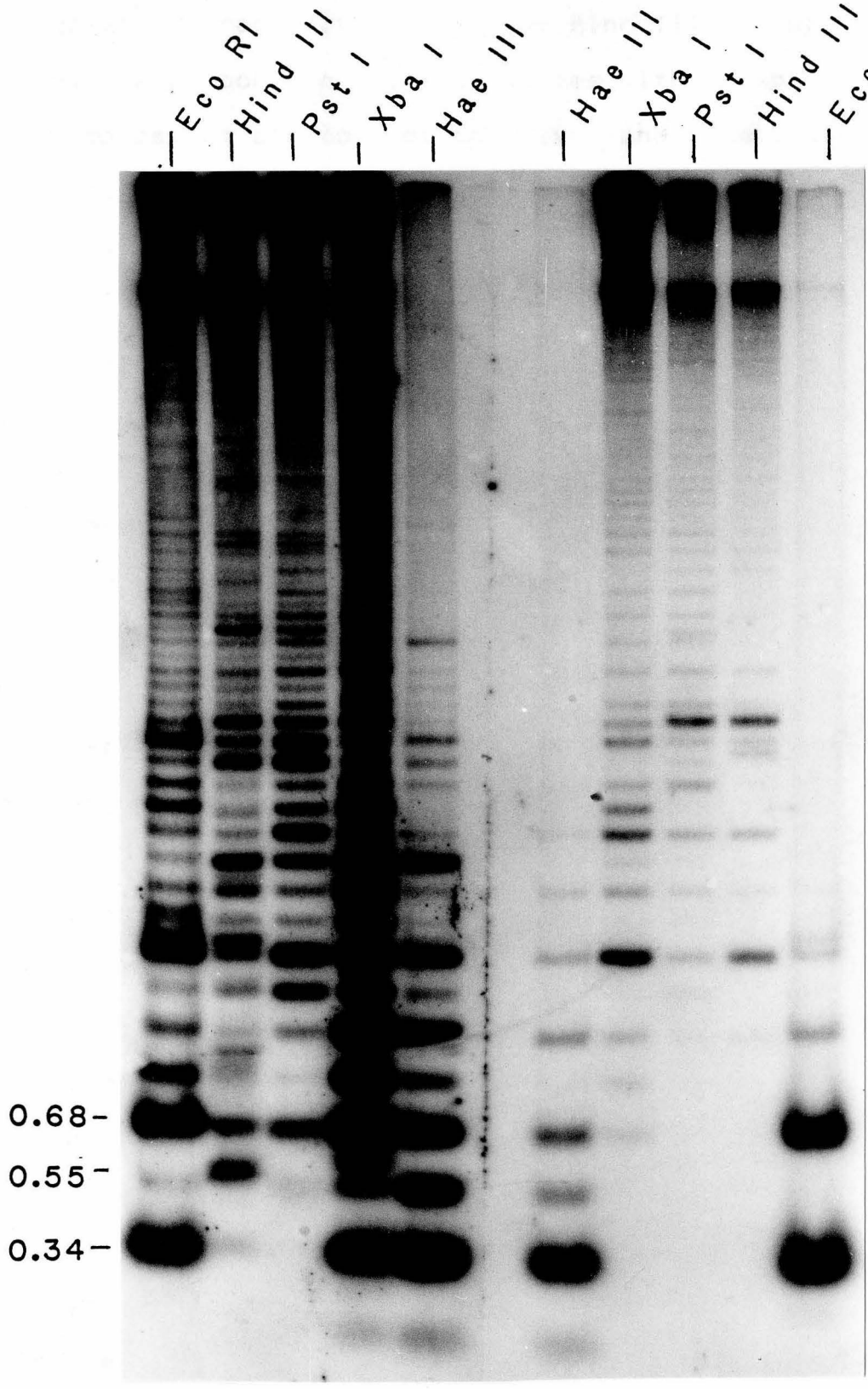

KpnI
1.8kb
KpnI
1.2kb
KpnI
1.5kb
KpnI
1.8kb
KpnI
5'
pHH550-2
3'

Figure 9. Genomic Blots probed with pHH550-31 or pHE340-64.
Placental DNA was digested with the restriction enzymes indicated, blotted on Gene Screen Plus and probed with pHH550-31 (the 5 left lanes) and pHE340-64 (the 5 right lanes). Sizes of selected fragments are indicated.


Eco RI
Hind III
Pst I
Xba I
Hae III
Hae III
Xba I
Pst I
Hind III
Eco RI
0.68
0.55
0.34

there is great heterogeneity within the Hind III 550 bp family. All lanes contain multimer series with an apparent 170 bp repeat and four of the five lanes contain high molecular weight DNA. This indicates that the basic repeating unit must have a number of sequence variants. The presence of multimer series means that a variant restriction site is located at analogous positions in all the repeats in which it occurs.

The 170 bp repeating unit is characteristic of alphoid-like sequences (Wu and Manuelidis, 1980). In order to verify that pHH550-31 is an alphoid sequence, identical digests were probed with either pHH550-31 or a known alphoid Eco RI 340 bp repeat (pHE340-64) (Figure 9). When the blots were lined up side-by-side, it was clearly visible that sequences that hybridize to pHH550-31 are composed of the characteristic alphoid 170 bp repeat. Dramatic differences are seen in the pattern of sequence variants for pHH550-31 and the known Eco RI 340 bp sequence (Figure 9). The sequences that hybridize to pHH550-31 show a great deal more heterogeneity than those sequences that hybridize to pHE340-64. For example in the comparable digests using Eco RI there is a much more extensive multimer series for the DNA probed with pHH550-31 than that which is probed with pHE340-64. The same is true with the other restriction digests. Among the sequence variants hybridizing to pHH550-31 are non-integral mem-

Figure 10. Restriction map of pHH550-31.
Locations of restriction sites for several enzymes are indicated along with the size scale. Restriction enzymes that not cut in this sequence are Ava I, Bam HI, Cla I, Eco RI, Hae III, Hpa II, Pst I, Pvu II, Sal I, Raq I, Xba I, and Xho I.

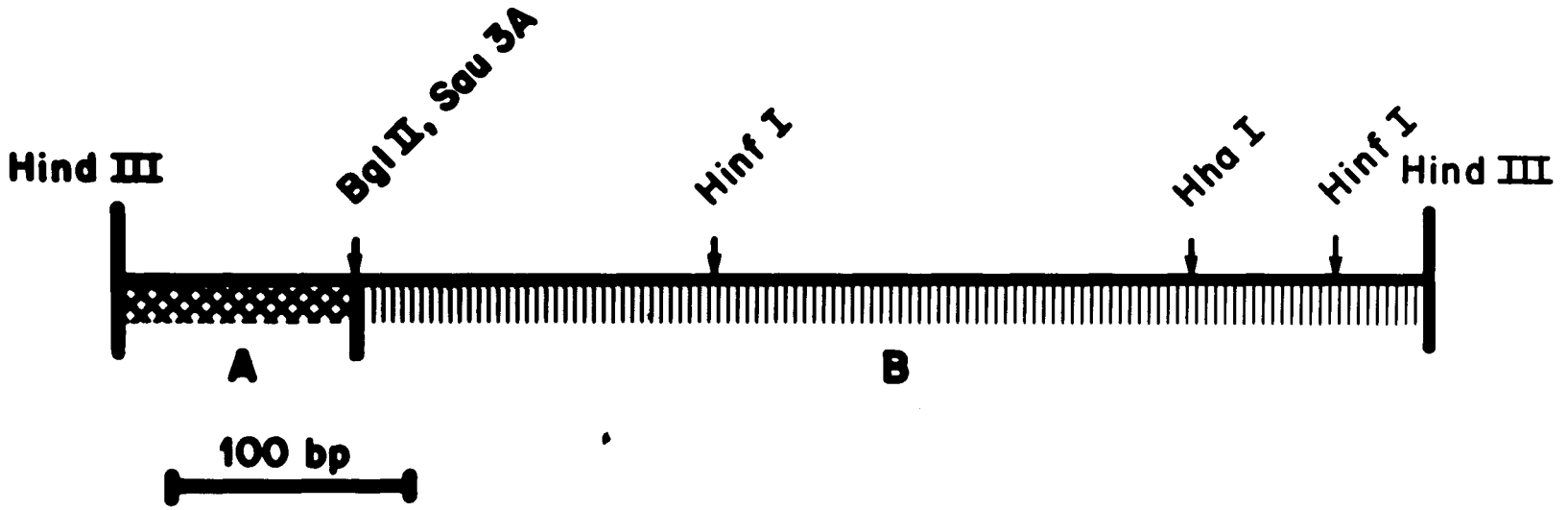

Hind III
Bgl II, Sau 3A
Hinf I
Hha I
Hinf I
Hind III
A
B
100 bp

Figure 11. Blot of pHE340-64 Probed with pHH550-31 at Varying Stringencies.
Each lane contains pHE340-64 digested with Eco RI and probed with pHH550-31. The hybrization of the left lane was done under standard stringency conditions (see Materials and Methods). The hybridization conditions used in the right lane were altered as follows: the second wash was done using 2X SSC at 65° C. and 1% SDS and the final wash was done using 0.1X SSC at room temperature. If cross-hybridization had occurred a band of 340 bp, the Eco RI 340 bp insert, would be seen.

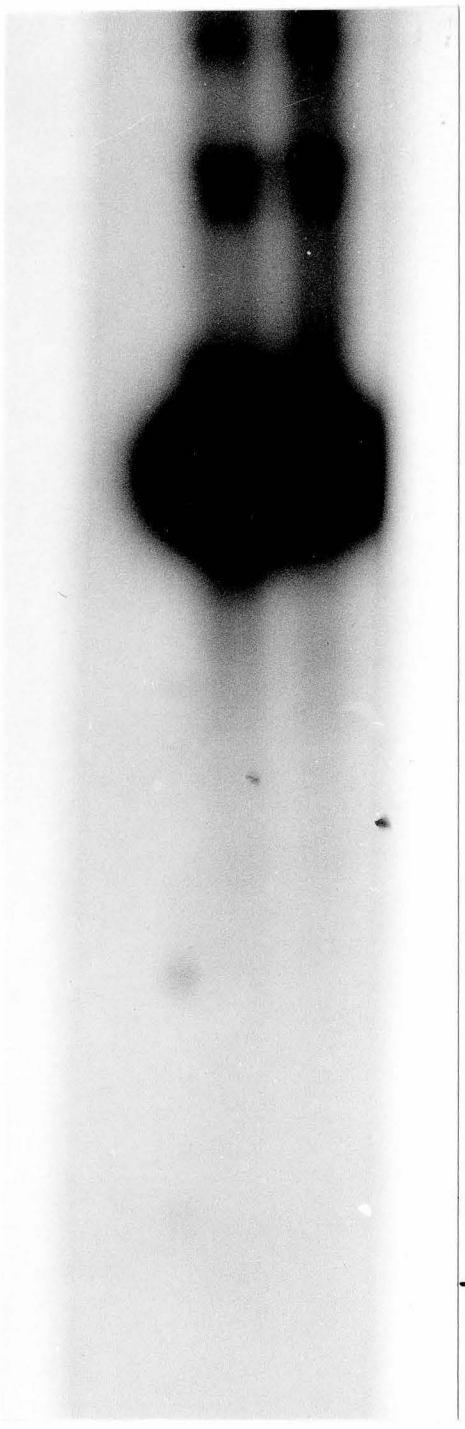
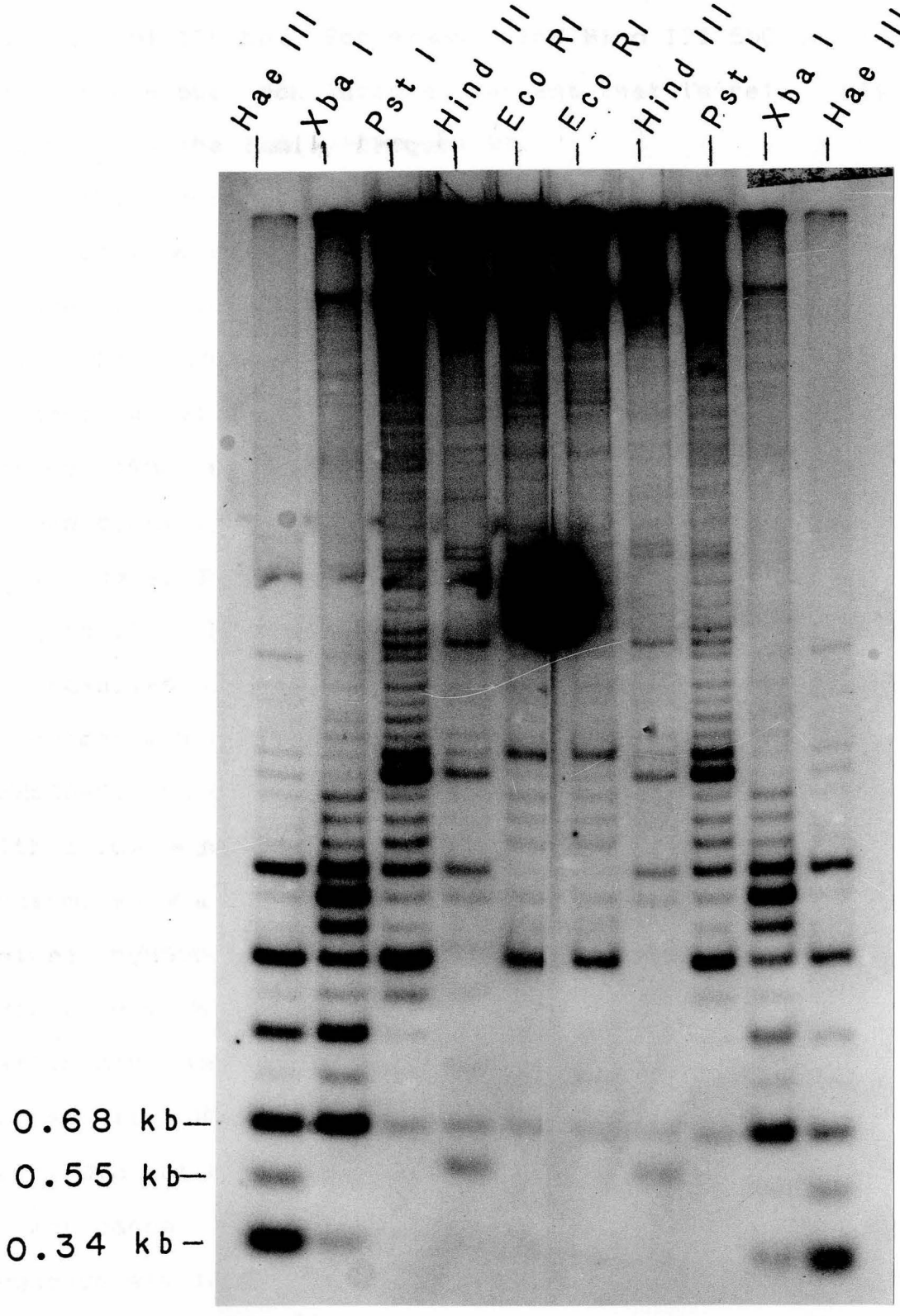

—0.34 kb

Figure 12. Blot of Genomic DNA Probed with pHH550-31 at Varying Stringencies.
Placental DNA was digested with the restriction enzymes indicated, blotted to Gene Screen Plus and probed with pHH550-31. The 5 left lanes were washed using standard stringencies (see Materials and Methods). The right lanes were subjected to the second wash of 2X SSC at 65° C. and the third wash using 0.1X SSC at room temperature. Sizes for selected fragments are indicated.


Hae III
Xba I
Pst I
Hind III
Eco RI
Eco RI
Hind III
Pst I
Xba I
Hae III
0.68 kb
0.55 kb
0.34 kb

bers. These are members whose sequence length is not a multiple of 170 bp. For example the Hind III 550 bp fragment is one such non-integral variant that is relatively uncommon in the family (Figure 9).

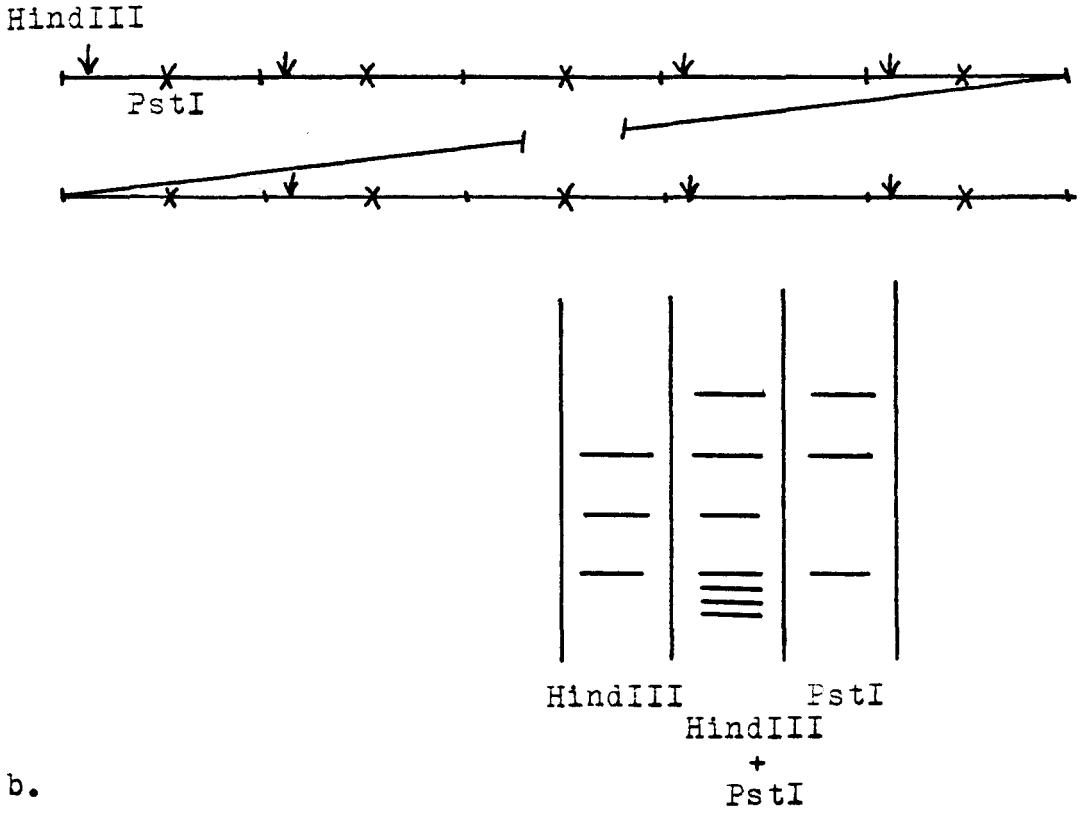
The restriction map for pHH550-31 (Figure 10) is distinct from other previously characterized alphoid sequences (Willard, 1985, Waye and Willard, 1985, Wolfe et al., 1985, Jorgensen et al., 1986, Jabs et al., 1984, Mitchell et al., 1985, Deville et al., 1985, Manuelidis and Wu, 1980, Gray et al., 1985). To determine if pHH550-31 was closely related to the Eco RI alphoid sequence, a known Eco RI 340 bp repeat was hybridized with pHH550-31 (Figure 11). Duplicate blots were probed at two different stringencies. In both cases no cross-hybridzation of the sequences was seen, thus supporting the hypothesis that pHH550-31 and Eco RI 340 are indeed distinct sequences with a low degree of sequence similarity. Next it was determined whether significant sequence similarity exists between pHH550-31 and the variant family members which hybridize with it. This was done by varying the hybridization stringencies on duplicate blots of genomic DNA probed with pHH550-31 (Figure 12). An increase in stringency did not cause the disappearance or loss of intensity of any bands. Thus, this serves as a verification of the sequence similarity that exists among members of the pHH550 bp family.

After seeing the heterogeneity that exists in the Hind III 550 bp family, it was important to know whether or not the numerous variants are randomly distributed throughout the family. In order to elucidate this structural organization it was necessary to perform side-by-side single and double digests of genomic DNA. If the variants recognized by one restriction enzyme are intermixed (Figure 13a) in the tandem arrays with variants recognized by other restriction enzymes then the restriction sites of one variant will be located within the repeat unit of another variant. In a side-by-side single and double digest the double digest will result in smaller restriction fragments. If the variants of one class (recognized by one restriction enzyme) are not intermixed with another class of variants then in side-by-side single and double digests the double digests will not contain additional smaller restriction fragments (Figure 13b). The double digest will contain all the fragments seen in both single digests. The blot (Figure 14) indicates that the members of each class tend to be located in distinct genomic domains apart from members of other classes. For example, a double digest with Eco RI and Hind III does not contain any fragments smaller than those seen in the single digests.

Owing to the chromosome specificity that many alphoid sequences exhibit (Willard, 1985), it was important to see

Figure 13. Organization of Alphoid Variants.
a. If the classes of alphoid variant sequences are randomly intermixed, then the restriction sites would also be intermixed. If genomic or chromosomal DNA were to be double digested, there would be additional, smaller fragments present that were not seen in the single digests. b. If each class of variant alphoid sequence is located in a distinct domain seperate from other variant classes then the restriction sites would remain in discrete regions, apart from each other. If genomic or chromosomal DNA were to be double digested, the additive effect would be seen, and there would be no additional fragments present that were not present in the single digests. The double digest would have all the bands found in both single digests.

a.


HindIII
PstI
HindIII
PstI
HindIII
+
PstI

b.

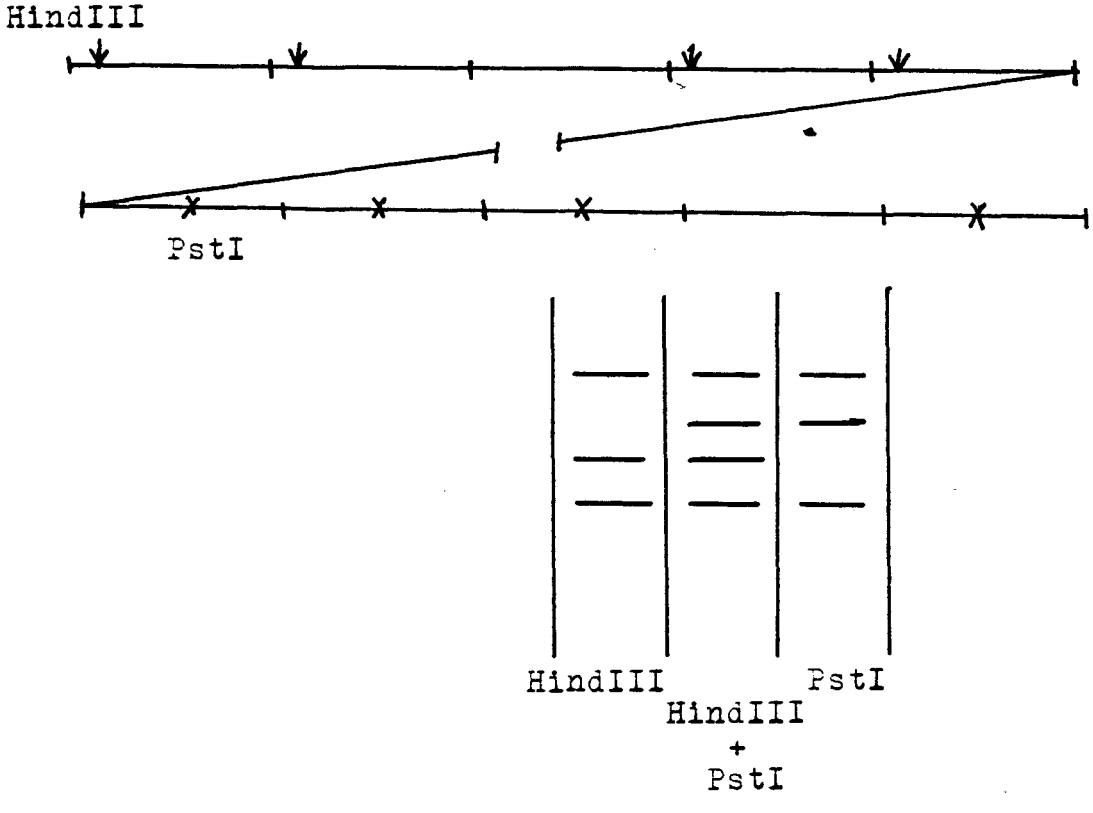
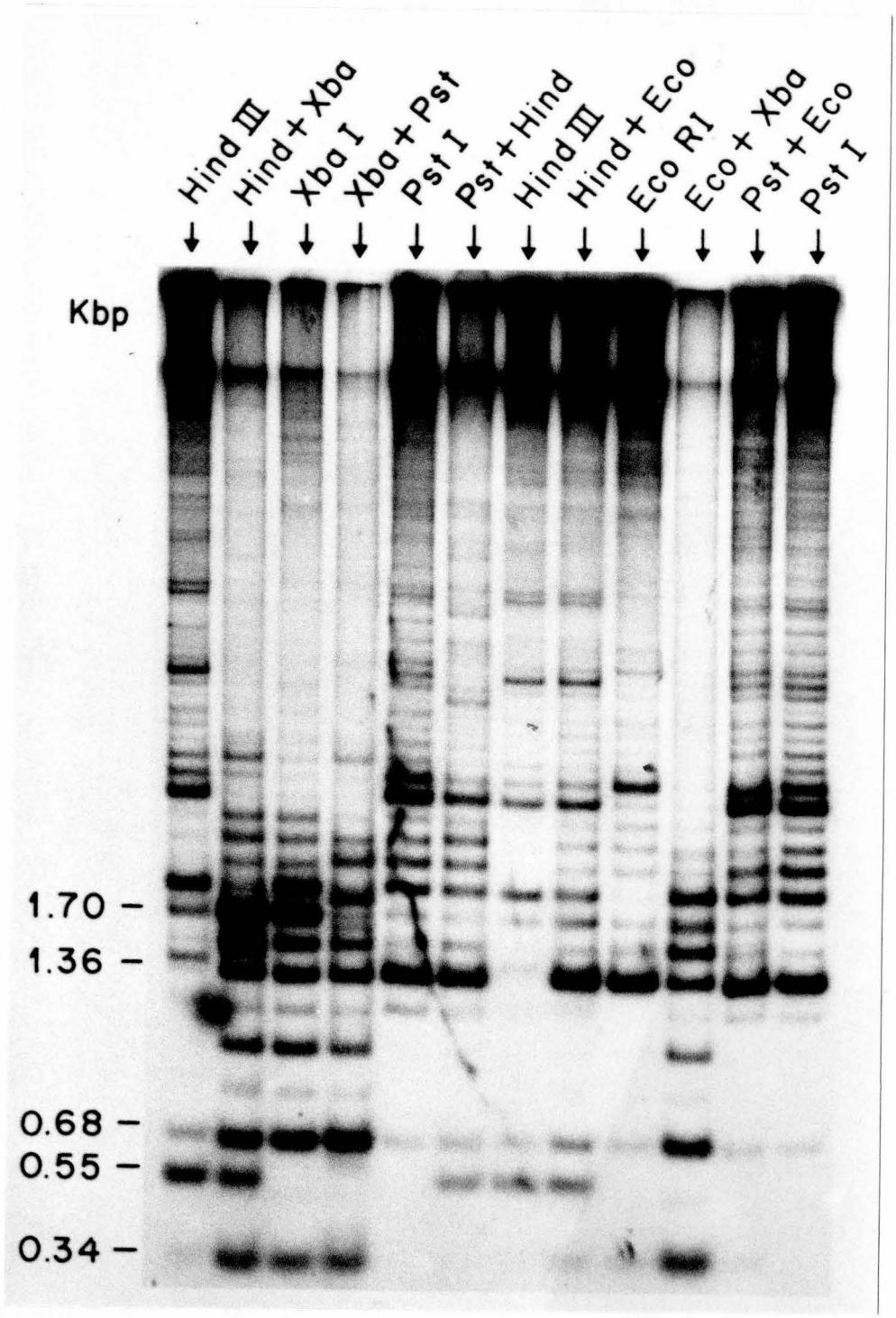

HindIII
PstI
HindIII
PstI
HindIII
+
PstI

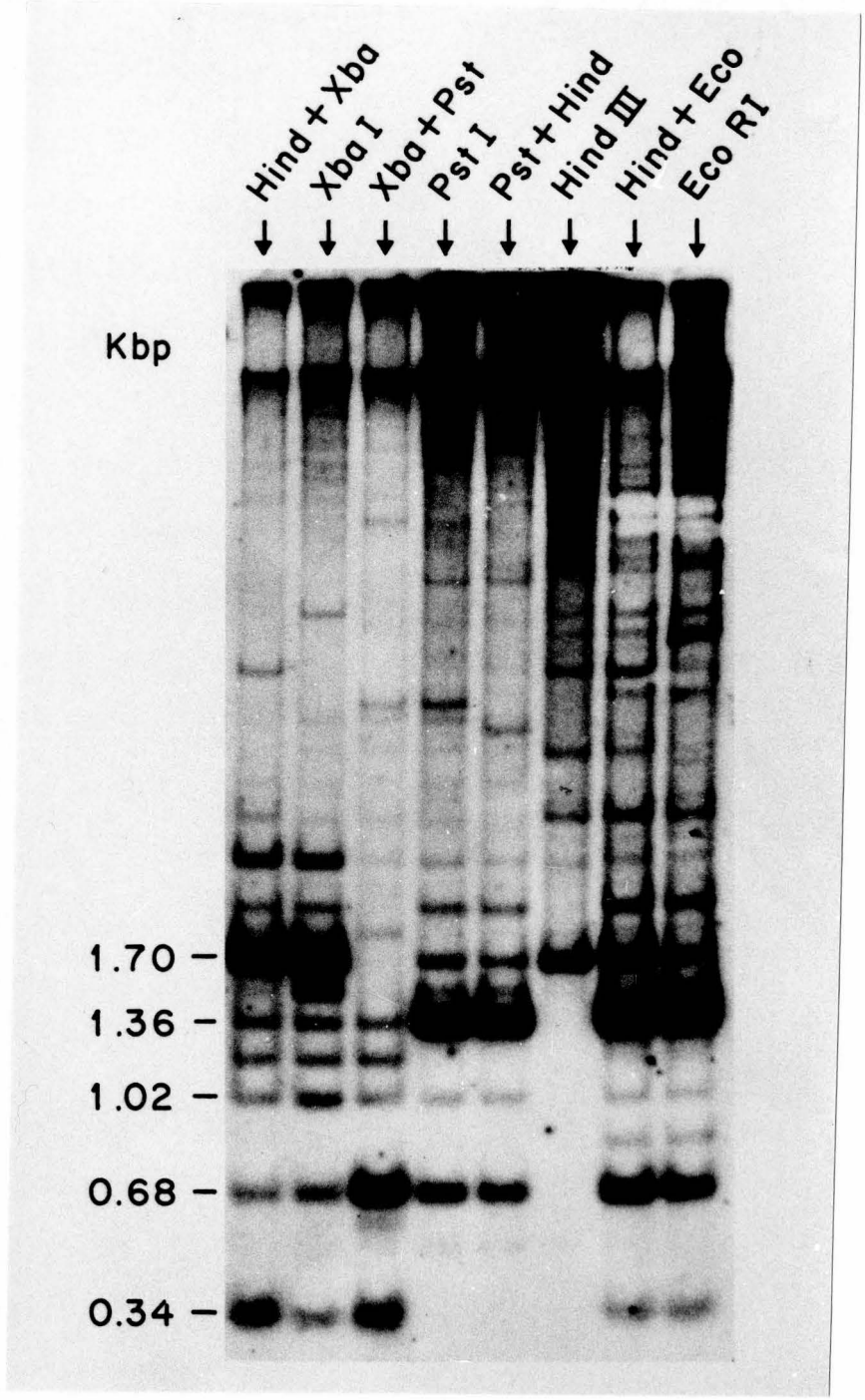
Figure 14. Double Digests of Genomic DNA Probed with pHH550-31.
Placental DNA was digested with the restriction enzymes indicated, blotted to Gene Screen Plus and probed with pHH550-31 as described in Materials and Methods. Sizes for selected fragments are indicated.


Hind III
Hind + Xba
Xba I
Xba + Pst
Pst I
Pst + Hind
Hind III
Hind + Eco
Eco RI
Eco + Xba
Pst + Eco
Pst I
Kbp
1.70 –
1.36 –
0.68 –
0.55 –
0.34 –

if such specificity was part of the character of the pHH550 family. In order to investigate the organization of the pHH550 family on a single chromosome, DNA from a hamster-human hybrid cell line containing only human chromosome 21 was singly and doubly digested and probed with pHH550-31 (Figure 15). On chromosome 21 multimer series that hybridize to pHH550-31 can be seen. There are variant members of the Hind III 550 bp family that are present in genomic DNA but missing on chromosome 21. The Pst I variant class has many higher molecular weight multimers that are present in genomic DNA but not found on chromosome 21. This is also true of the Hind III variant class. It is important to note that the Hind III 550 bp fragment present on genomic DNA (Figure 14) is absent on chromosome 21: this implies chromosome-specificity.

The next step was to see if on chromosome 21 the Hind III 550 bp family is found in distinct domains, as it is in genomic DNA, or whether the variant classes are intermixed. To do this it is necessary to compare single and double digests of chromosome 21 DNA (Figure 15). The additive effect is seen here, as it was in genomic DNA. Doubly digesting chomosome 21 DNA, for example with Hind III and Eco RI, does not yield any additional fragments (Figure 15). All bands present in each single digest are also present in the double digest (Figure 15). Thus even on a single chromosome members of a variant class tend to

Figure 15. Double Digests of Chromosome 21 DNA Probed with pHH550-31.
DNA from hybrid cell line 153-E9A was digested with the restriction enzymes indicated, blotted to Gene Screen Plus and probed with pHH550-31 as described in Materials and Methods. Sizes for selected fragments are indicated.


Hind + Xba
Xba I
Xba + Pst
Pst I
Pst + Hind
Hind III
Hind + Eco
Eco RI
Kbp
1.70
1.36
1.02
0.68
0.34

be located in domains apart from members of other classes.

The sequence of pHH550-31 was obtained using fragments that were labelled at the 5′ end. The following sequencing strategy (Figure 16) was used: 1) The Eco RI site, outside the insert, was labelled and 285 bp were read from the left end of the insert toward the right. 2) Next the Bgl II site was labelled and 312 bp were read from that site toward the right. This created an overlap of 110 bp with the sequence read from the the Eco RI site. 3) Finally, the right hand Hind III site was labelled and 220 bp of sequence was read on the opposite strand back toward the left. This created an overlap of 90 bp with the sequence read from the Bgl II site.

The overall sequence length of pHH550-31 (Figure 17) is 551 bp. Since pHH550-31 is an alphoid sequence it should contain internal repeats of about 170 bp in length. The sequence can be arranged as three and one quarter repeating units, each of a distinctive size. The first repeat unit (I) is 171 bp, II is 167 bp, III is 170 bp and IV is only 42 bp of what appears to be a 168 bp repeat unit (Figure 18). There is a deletion at position 99 to permit maximum sequence similarity between repeat units. In Figure 18 the alignment of the repeat units places a Hind III or a modified Hind III site at the beginning of each repeat unit. The sequence similarity between I and II is 63.2%, between II and III 62.4% and between I and

Figure 16. Sequencing Strategy for pHH550-31.
The arrows indicate the direction and length of sequencing from a particular restriction site. See text for detailed sequencing strategy.

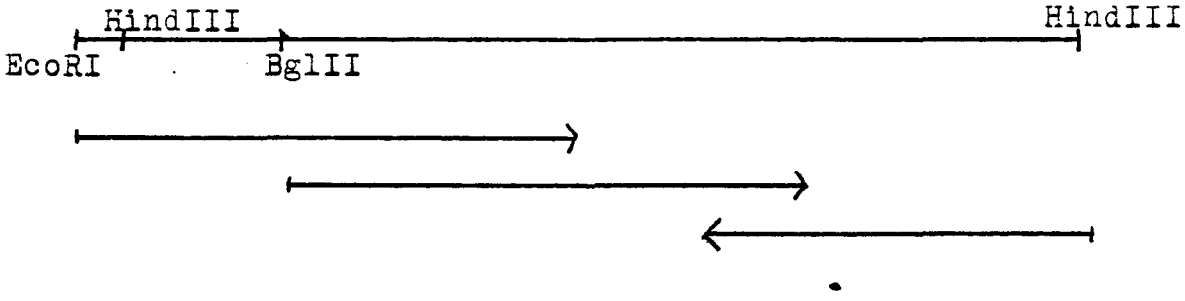

HindIII
HindIII
EcoRI
BglII

Figure 17. Sequence of pHH550-31.
The nucleotide sequence of pHH550-31 is shown with the restriction sites indicated on the map in Figure 10.

```
    *    *    *    *    *    *    *    *    *    *
    5   10   15   20   25   30   35   40   45   50
AAGCTTCATTGGGATGTTTTAATTGAAGTCACAGTGTTGAACAGTCCCTT
HindIII

    *    *    *    *    *    *    *    *    *    *
   55   60   65   70   75   80   85   90   95  100
TCATAGAGCAGGTTTGAAACACTCTTTTTGTAGTATCTGGAAGTGGACAT

    *    *    *    *    *    *    *    *    *    *
  105  110  115  120  125  130  135  140  145  150
TTGGAGAGATCTCAGGAATACGGTGTTAAAGGAAATATCTTCCAATAAAA
       BglII

    *    *    *    *    *    *    *    *    *    *
  155  160  165  170  175  180  185  190  195  200
GCTAGATAGAAGCAATGTCAGAAACTTTTTCATGATGTATCTACTCAGCT

    *    *    *    *    *    *    *    *    *    *
  205  210  215  220  225  230  235  240  245  250
AACAGAGTTGAACCTTTCTTTTGAGAGAGCAGTTTTGAAACACTCTTTTT

    *    *    *    *    *    *    *    *    *    *
  255  260  265  270  275  280  285  290  295  300
GTGGAATCTGCAAGTGGATATTTGTCTAGCTTTGAGGATTTCGTTGGAAA
   HinfI

    *    *    *    *    *    *    *    *    *    *
  305  310  315  320  325  330  335  340  345  350
CGGGATTACATATAAAAAGCAGACAGCAGCATTCCCAGAAACTTCTTTGT

    *    *    *    *    *    *    *    *    *    *
  355  360  365  370  375  380  385  390  395  400
GAAGTTTGCATTCAAGTCACAGAGTTGAACATTCCCTTTCATAGAGCAGG

    *    *    *    *    *    *    *    *    *    *
  405  410  415  420  425  430  435  440  445  450
TTTGAAACACTCTTTTTGTAGTATCTGGATGTGGACTTTGGTGCGCTTTC
                                          HhaI

    *    *    *    *    *    *    *    *    *    *
  455  460  465  470  475  480  485  490  495  500
AGGCGTATGGTGAAAAAGGAAATATCTTCCCCTGAAAACTAGACAGAAGC

    *    *    *    *    *    *    *    *    *    *
  505  510  515  520  525  530  535  540  545  550
ATTCTCAGAATCTTATTTGTGATATGCGCTCTCAACTAACAGTGTTGAAGC
       HinfI                                  HindIII
```

Figure 18. Comparison of the Repeating Units within pHH550-31.
The repeating units, I, II, III, and IV are shown. The number to the right of the Roman numeral is that repeat's starting point within the full length sequence. A highly conserved region exists between positions 49 and 103 in each repeat unit.

```
                 *    *    *    *    *    *    *    *    *    *
                 5   10   15   20   25   30   35   40   45   50
I                AAGCTTCATTGGGATGTTTTAATTGAAGTCACAGTGTTGAACAGTCCCTT
II 172           AAACTTTTTCATGATGTATCTACTCAGCTAACAGAGTTGAACCTTTCTTT
III 339          AAACTTCTTTGTGAAGTTTGCATTCAAGTCACAGAGTTGAACATTCCCTT
IV 508           AATCTTATTTGTGATATGCGCTCTCAACTAACAGTGTTGAAGC

                 *    *    *    *    *    *    *    *    *    *
                55   60   65   70   75   80   85   90   95  100
I                TCATAGAGCAGGTTTGAAACACTCTTTTTGTAGTATCTGGAAGTGGACAT
II               TGAGAGAGCAGTTTTGAAACACTCTTTTTGTGGAATCTGCAAGTGGATAT
III              TCATAGAGCAGGTTTGAAACACTCTTTTTGTAGTATCTGGATGTGGACXT

                 *    *    *    *    *    *    *    *    *    *
               105  110  115  120  125  130  135  140  145  150
I                TTGGAGAGATCTCAGGAATACGGTGTTAAAGGAAATATCTTCCAATAAAA
II               TTGTCTAGCTTTGAGGATTTCGTTGGAAACGGGATTACATATAAAAAGCA
III              TTGGTGCGCTTTCAGGCGTATGGTGAAAAAGGAAATATCTTCCCCTGAAA

                 *    *    *    *
               155  160  165  170
I                GCTAGATAGAAGCAATGTCAG
II               GACAGCAGCATTCCCAG
III              ACTAGACAGAAGCATTCTCAG
```

III 84.8%. Since there is more sequence similarity between I and III than between I and II or II and III, it may be that pHH550-31 has a basic 340 bp repeat. It is thus possible that the overall repeating unit is 338 bp. Comparing the first 42 bp of II with IV a sequence similarity of 69% is seen. This cannot be considered an accurate estimate of an overall comparison of II and IV because the region of greatest similarity between the separate repeating units is 3′ to the first 42 bp. The two subunits of the Eco RI 340 bp fragment have a sequence similarity of 73% (Wu and Manuelidis, 1980) as compared with 63.2% for I and II of pHH550-31 and 62.4% for II and III of pHH550-31. Considering only the full-length subunits of pHH550-31, they have an average sequence similarity of 70.1%. This can be compared with a less than 68% average sequence similarity for the subunits within the Xba I 682 bp sequence (Gray et al., 1985). The sequence similarity between subunits of pHH550-31 is less than that seen within p82H (Mitchell et al., 1985) where the 14 subunits have a sequence similarity of 74-92%. When comparing the repeat units of pHH550-31 there are five regions of 4-19 bp that are completely conserved: positions 36-41, 63-81, 85-89, 93-97 and 100-103 (Figure 19). A highly conserved region of 86% sequence similarity (positions 47-103) is found in the middle of each repeating unit (Figure 18).

In order to verify that pHH550-31 is a unique alphoid family a sequence comparison of pHH550-31 with other alphoid families was done. The analogous regions of pHH550-31 and the Eco RI 340 bp alphoid sequence (Manuelidis and Wu, 1980) have a sequence similarity of 73.9%, while the analogous regions of the Xba I 682 bp alphoid sequence (Gray et al., 1985) and pHH550-31 share a sequence similarity of 82.9% (Figure 19). Thus these comparisons provide verification that pHH550-31 represents a unique alphoid family. There are short regions that are completely conserved between the three alphoid sequences: for example, positions 64-77, 130,-142, 257-272 and 472-484 (Figure 19). Such regions might have been conserved to permit binding of a specific protein. Outside of these conserved regions the mismatches seen between the three alphoid sequences are randomly distributed.

Alphoid sequences in African green monkey (AGM) are known to bind a specific protein thought to control nucleosome phasing (Strauss and Varshavsky, 1984). When comparing pHH550-31 with the alpha satellite of AGM used by Strauss and Varshavsky (1984), the comparable sites for the attachment of "alpha protein" would be at positions 19-25, 96-102 and 133-139 of pHH550-31 (Figure 18). A 42.8% overall mismatch is seen when comparing all the subunits of pHH550-31 (Figure 18) with the first binding site (5′-TTAATTC-3′) of the AGM alpha component DNA

Figure 19. Comparison of Eco RI 340 bp, pHH550-31 and Xba I 682 bp Alhpoid Sequences.
The top line is the Eco RI 340 bp sequence (Maunelidis and Wu, 1980), beginning at position 9. The second line is the Xba I 682 bp (Gray et al., 1985) sequence beginning at its analogous position to the Eco RI 340 bp sequence. The third line is pHH550-31. There are fully conserved regions between all three sequences, for example at positions 64-77, 130-142, 257-272 and 472-484.

```
                 *    *    *    *    *    *    *    *    *    *
                 5   10   15   20   25   30   35   40   45   50
I          TAACTTCCTTGTGTTGTGTGTATTCAACTCACAGAGTTGAACGATCCTTT
Xba I      GAACTTCTTTGTGATGTTTGCATTCACGTCACAGAACTGAACATTCCCTT

-31        AAGCTTCATTGGGATGTTTTAATTGAAGTCACAGTGTTGAACAGTCCCTT


                 *    *    *    *    *    *    *    *    *    *
                55   60   65   70   75   80   85   90   95  100
I          ACACAGAGCAGACTTGAAACACTCTTTTTGTGGAATTTGCAAGTGGAGAT
Xba I      TCATAGAGCATGTTTGAAACACTCTTTCTGTAGTATCTACAAACGGACAT

-31        TCATAGAGCAGGTTTGAAACACTCTTTTTGTAGTATCTGGAAGTGGACAT


                 *    *    *    *    *    *    *    *    *    *
               105  110  115  120  125  130  135  140  145  150
I          TTCAGCCGCTTTGAGGTCAATGGTAGAATAGGAAATATCTTCCTATAGAA
Xba I      TTCAAACGCTTTCAGGCCTATGGTGAGAAAGGAAATATCTTCAAATAAAA

-31        TTGGAGAGATCTCAGGAATACGGTGTTAAAGGAAATATCTTCCAATAAAA


                 *    *    *    *    *    *    *    *    *    *
               155  160  165  170  175  180  185  190  195  200
I-II       ACTAGACAGAATGATTCTCAGAAACTCCTTTGTGATGTGTGCGTTCAACT
Xba I      ACTAGACAGAAGCATTCTCAGAAACTTATTTGCGATGTGTGTCCTCAACT

-31        GCTAGATAGAAGCAATGTCAGAAACTTTTTCATGATGTATCTACTCAGCT


                 *    *    *    *    *    *    *    *    *    *
               205  210  215  220  225  230  235  240  245  250
II         CACAGAGTTTAACCTTTCTTTTCATAGAGCAGTTAGGAAACACTCTGTTT
Xba I      AACAGAGTTGAACCTTTCTTTTGATACAACATTTTGGAAACACTCTTTTT

-31        AACAGAGTTGAACCTTTCTTTTGAGAGAGCAGTTTTGAAACACTCTTTTT


                 *    *    *    *    *    *    *    *    *    *
               255  260  265  270  275  280  285  290  295  300
II         GTAAAGTCTGCAAGTGGATATTCAGACCTCTTTGAGGCCTTCGTTGGAAA
Xba I      GTAGAATCTGCAAGTGGATATTTGAATAGCTTTGAAGGTTTCGTTGGAAX

-31        GTGGAATCTGCAAGTGGATATTTGTCTAGCTTTGAGGATTTCGTTGGAAA
```

```
                  *    *    *    *    *    *    *    *    *    *
                305  310  315  320  325  330  335  340  345  350
II-I       CGGGATTTCTTCATATTATGXXCTAGACAGAAGAATTCTCAGTAACTTCC
Xba I      CGGGAATATCTTCATATAAAATCAAGACAGAAGCATTCTCAGAAACTTCT

-31        CGGGATTAXCATXATAAAAAGXCAXGACAGCAGCATTCCCAGAAACTTCT


                  *    *    *    *    *    *    *    *    *    *
                355  360  365  370  375  380  385  390  395  400
I          TTGTGTTGTGTGTATTCAACTCACAGAGTTGAACGATCCTTTACACAGAG
Xba I      CTGTGATGTTTGCATTCAACTCATAGAGTTGAACACTTCCCTTCGTACAG

-31        TTGTGAAGTTTGCATTCAAGTCACAGAGTTGAACATTCCCTTTCATAGAG


                  *    *    *    *    *    *    *    *    *    *
                405  410  415  420  425  430  435  440  445  450
I          CAGACTTGAAACACTCTTTTTGTGGAATTTGCAAGTGGAGATTTCAGCCG
Xba I      CAGGTTTGAAACACTCTTTTTGTAACATTTGGAAGTGGACATTTGCAGCG

-31        CAGGTTTGAAACACTCTTTTTGTAGTATCTGGATGTGGAXCTTTGGTGCG


                  *    *    *    *    *    *    *    *    *    *
                455  460  465  470  475  480  485  490  495  500
I          CTTTGAGGTCAATGGTAGAATAGGAAATATCTTCCTATAGAAACTAGACA
Xba I      CTTTGAGGCCTATGTTGAAAAAGGAAATATCTTCTCCTAAAAACCAGACA

-31        CTTTCAGGCGTATGGTGAAAAAGGAAATATCTTCCCCTGAAAACTAGACA


                  *    *    *    *    *    *    *    *    *    *
                505  510  515  520  525  530  535  540  545  550
I-II       GAATCATTCTCAGAAACTCCTTTGTGATGTGTGCGTTCAACTCACAGAGT
Xba I      GAAGCATTCTCAGAAACTTCCTTGTGATGTGTGTACTCGAGTAACAGAGT

-31        GAAGCATTCTCAGAATCTTATTTGTGATATGCGCTCTCAACTAACAGTGT


                  *
                555
II         TTAACC
Xba I      TGAACC

-31        TGAAGC
```

(Strauss and Varshavsky, 1984). The second comparable binding site (5′-TTTATAG-3′) (Strauss and Varshavsky, 1984) has an overall average mismatch of 76.2% with its analogous sites of all pHH550-31 subunits (Figure 18). In a comparison of the third site (5′-AAATATC-3′) (Strauss and Varshavsky, 1984) with all its pHH550-31 counterparts the average mismatch is 19%. Between the individual repeats of pHH550-31 there is great divergence in the extent of sequence similarity to AGM alpha protein binding sites. Some show a good match and for others it is poor, thus making the binding of an alpha protein to pHH550-31 unlikely.

DISCUSSION

This study has characterized two newly-discovered repetitive sequences of human DNA, one a dispersed repeat and the other a tandem repeat. The dispersed sequence is a variant of a previously described interspersed repetitive family, L1Hs. A new alphoid family is represented by the tandemly repeated sequence. Their characterization has added to the rapidly expanding volume of information about the genomic organization of the reiterated portion of the human genome. Comparison of these newly discovered sequences to previously characterized ones indicate that previous concepts concerning sequence homogenization of repetitive sequence families may need to be revised.

A Dispersed Variant

All evidence indicates that pHH550-2 is a variant member of the L1Hs family. The strong sequence similarity, 92.3%, seen between pHH550-2 and T-beta-G41 (a characterized L1Hs member) provides strong evidence that the two sequences are members of the same family. In addition previous work showed the hybridization of pHH550-2 to a 1.9 kb Pst I fragment and Kpn I fragments of 0.6 kb and 0.9 kb (Doering and Burket, unpublished data). T-beta-G41 has similar fragments. The majority of sequences containing pHH550-2 are found as high molecular weight DNA.

The fact that a variant occurs at the 5′ end of L1Hs may not be surprising since this is the end at which the truncations normally occur (Hwu et. al., 1986). However, it is surprising to see such a high degree of sequence similarity (92.3%) between two interspersed sequences of the same family, pHH550-2 and T-beta-G41. The method of amplification and dispersal of interspersed sequences is thought to contribute to a lack of sequence similarity (Orgel and Crick, 1980). The region of high sequence similarity that is seen in the three L1Hs family members shown in Figure 7 is found in the 5′ 1.8 kb subregion of T-beta-G41 (Figure 8). Between positions 26-219 of pHH550-2 the sequence is highly conserved (Figure 7). This region of conservation is not located within any open reading frame. Therefore it may have some as yet unknown function.

Conservation of a DNA sequence is thought to be indicative of importance, either functional or structural. Some believe that L1Hs may still have members that are transcriptionally active (Fujita et. al., 1987). Thus, these 200 base pairs may be maintained as a significant segment of a control mechanism. In the comparison of analogous regions between Kpn A and pHH550-2 (Figure 7) the existence of a region of conservation adjacent high sequence dissimilarity is due to the rearrangement within Kpn A. The permutation and the deletion of Kpn A (Potter, 1984) occur 3′ of position 220 causing a lack of sequence

analogy between pHH550-2 and Kpn A.

The sequence differences seen between pHH550-2, Kpn A and T-beta-G41 in the analogous regions could easily be point mutations that have occurred during the evolution of the sequence, since the majority of the changes are only single base alterations. There are a few places where two adjacent bases have been affected.

Future work with pHH550-2 should involve studies of its chromosomal location, since location may have a relationship to function. Is it located on specific chromosomes or subregions of the chromosomes? Does it show a preference as to its location near functional and/or processed genes? If so is it because the sequence is a control for transcriptional acitivity? In the future it will be desirable to compare additional L1Hs family members to see if they too share a high degree of sequence similarity with each other and those previously described family members. If so, then this could be a stronger indication of some functional significance for this sequence.

An Alphoid Variant

The characterization of pHH550-31 has proven it to be yet another new family of alphoid sequences. Although all alphoid families have repeating units of 170 bp they may have a sequence similarity as limited as 50% (Mitchell et al., 1985). This divergence is exhibited by the lack of cross-hybridization of pHH550-31 with a known Eco RI 340

bp repeat (Figure 11).

Chromosome-specificity is characteristic of some of the alphoid families (Waye and Willard, 1985, Willard et al., 1986, Tyler-Smith and Brown, 1987) including the Hind 550 bp alphoid family. This study showed a number of sequence variants of this family are present in the genome with the exception of chromosome 21. The Hind III 550 bp sequence is not found on chromosome 21, but sequences similar to it are located there.

Structural organization within the genome and on a single chromosome is similar. Variant classes of pHH550-31, each recognized by a specific restriction enzyme, are located in individually distinct domains. The domains are not intermixed and more than one variant class domain can be present on a single chromosome. It is known that the Eco RI 340 bp family is also organized in this manner (Doering et al., 1986). Thus, this may be a feature of organization for all alphoid families and may have functional significance.

Hitherto, the existence of multiple alphoid families on a single chromosome has not been realized. The Eco RI 340 bp family (Jorgenson et al., 1987), the Xba I 682 bp family (Gray et al., 1985) and variant members of the Hind III 550 bp family are found on chromosome 21. Is the presence of multiple alphoid families on a single chromosome crucial to chromosomal structure and/or function? It

is possible that members of each alphoid family are present on each and every chromosome and must be there for functional purposes. It is not yet known if members of individual alphoid families on a single chromosome are all clustered together or intermixed with other alphoid families.

The basic repeating unit of pHH550-31 is approximately 170 bp, but since the first repeating unit (I) and the third (III) show much greater sequence similarity (84.8%) than I and II (63.2%) or II and III (62.4%) it is highly possible that the repeat is 338 bp. A comparison of II and IV is insignificant because the region of high sequence similarity within the repeating units is beyond the 42 bp contained in IV of pHH550-31. The possibility also exists that there are four or more distinct subunits that form a higher-order repeat. In order to confirm this it would be necessary to sequence a longer variant Hind III member.

There is a region of high sequence similarity in the middle of each repeating unit, positions 47-103 (Figure 16). As yet there is no known function for this region of conservation. There are smaller completely conserved regions, positions 36-41, 55-61, 63-81, 85-89, 93-96 and 100-103 (Figure 16). Nuclear binding proteins associate with short conserved sequences of alpha component of AGM and could function in nucleosome phasing. Therefore, the possibility exists that the above conserved sequences

could be protein binding sites which could function in nucleosome phasing. However, after a comparison of the sequences, there is not enough similarity between the appropriate areas of pHH550-31 and AGM alpha satellite (Strauss and Varshavsky, 1984) to allow such a protein to bind to pHH550-31. The possibility exists that some alphoid families will bind alpha proteins, whereas other will not.

Sequence similarity can be used as an indicator of the evolutionary age of a given sequence. The longer a sequence family has been present in a genome the longer it has been exposed to mutational forces, thus the more divergent the older the sequence. The degree of sequence similarity seen between adjacent subunits of pHH550-31 is significantly lower than the 73% found between the Eco RI 340 bp family subunits (Wu and Manuelidis, 1980) and the 74-92% between the p82H alhpoid family subunits (Mitchell et al., 1985). Thus, pHH550-31 may have been undergoing sequence alteration for longer periods of time than some other alphoid families or may be under the control of some other homogenizing mechanism. A sequence comparison of the Eco RI 340 bp alphoid family (Wu and Manuelidis, 1980) and the Xba 682 family (Gray et. al., 1985) with pHH550-31 (Figure 19) shows that the sequences themselves are quite distinct, showing sequence similarities of 73.9% and 82.9% respectively. It has been established that the Eco RI 340

bp family is evolutionarily the oldest alphoid family (Wu and Manuelidis, 1980, Gray et al., 1985). The sequence similarity between the Eco RI tetramer and the Xba 682 bp sequence shows a sequence similarity of 78.2% (Gray et al., 1985). On this basis it can be stated that the Hind III 550 bp sequence diverged from the Eco RI 340 bp sequence before the Xba I 682 bp sequence.

In most cases, before function can be determined the detailed structure and organization of DNA sequences must be determined. Future work should look for additional alphoid sequences, the degree to which their subunits diverge from one another and the degree of divergence from other alphoid sequence families. It will be beneficial to compare conserved regions between subunits and between the various alphoid families in attempts to establish any functional aspects of their structure. It will be of interest to know if the variant sequence classes of other alphoid families are intermixed within themselves and/or with other alphoid families or are they found in distinct domains as is the Hind III 550 bp family. It will also be important to check whether variant members of all alphoid families are present on all chromosomes.

Alphoid sequences are preferentially located in the centromere region (Manuelidis, 1978b) and in addition some may be chromosome-specific. It is possible that these alphoid areas of heterochromatin whose domains do not

intermix serve to protect much smaller sequences of euchromatin. The euchromatin could possibly serve as an area of attachment of the spindle fibers of mitosis and meiosis. If the sequence of attachment is comparatively small and must remain unaltered, it is possible that proteins of a nature different from that of the "alpha protein" (Strauss and Varshavsky, 1984) could bind to the repetitive sequences surrounding it. This could possibly then afford some physical protection from mutational forces for the euchromatin. It is possible that early in its evolutionary history when the first single alphoid sequence began to amplify, an area of unequal crossing-over occurred that then encased the length of DNA that served as the point of attachment. Then the surrounding alphoid sequences could continue to amplify and undergo sequence divergence.

Conclusion

The two sequences, pHH550-2, the interspersed repeat and pHH550-31, the tandem repeat, appear to be contradictions of accepted concepts for evolution of the different forms of repetitive sequences. It is thought that dispersed sequences should show more dissimilarity within a family than tandemly repeated sequence families because sequences dispersed around the genome would not come under the homogenizing mechanisms as they are presently proposed (Smith. 1976). Thus individually the family members could

intermix serve to protect much smaller sequences of euchromatin. The euchromatin could possibly serve as an area of attachment of the spindle fibers of mitosis and meiosis. If the sequence of attachment is comparatively small and must remain unaltered, it is possible that proteins of a nature different from that of the "alpha protein" (Strauss and Varshavsky, 1984) could bind to the repetitive sequences surrounding it. This could possibly then afford some physical protection from mutational forces for the euchromatin. It is possible that early in its evolutionary history when the first single alphoid sequence began to amplify, an area of unequal crossing-over occurred that then encased the length of DNA that served as the point of attachment. Then the surrounding alphoid sequences could continue to amplify and undergo sequence divergence.

Conclusion

The two sequences, pHH550-2, the interspersed repeat and pHH550-31, the tandem repeat, appear to be contradictions of accepted concepts for evolution of the different forms of repetitive sequences. It is thought that dispersed sequences should show more dissimilarity within a family than tandemly repeated sequence families because sequences dispersed around the genome would not come under the homogenizing mechanisms as they are presently proposed (Smith. 1976). Thus individually the family members could

come under a greater variety of sequence altering influences. Given the present finding of high sequence similarity within the L1Hs family, some other mechanism of homogenization for dispersed sequences must exist that is presently unknown. Possibly the interspersed sequences are moving around the genome at a rate faster than that of mutation. The accumulation of tandemly repeated sequences occurs through lateral amplification, thus the repeated units remain linked to each other. Therefore, it is thought that they should have high sequence similarity. The homogenization process, unequal crossing over as proposed by Smith (1976), is thought to cause tandemly repeating sequences to have a high degree of sequence similarity. Through sequencing, the opposite has been shown to be true, at least in this case: the tandemly repeating sequence, pHH550-31, is more diverse in sequence from other alphoids than the interspersed sequence, pHH550-2, is from other members of its interspersed family, L1Hs. Possibly the homogenizing mechanisms of the tandem repeats are not as efficient as first thought, or for as yet some unknown reason, there is a functional requirement for heterogeneity to be maintained in alphoid sequences. If the results seen here prove to be the rule and not the exception, then a major revision of concepts of how repetitive sequences evolve and maintain their sequence similarity must be sought.

REFERENCES

Blin, N. and Stafford, D.W. (1976). A general method for isolation of high molecular weight DNA from eukaryotes. Nuc.Acids Res. 3, 2303-2308.

Boeke, J.D., Garfinkel, D.J., Styles, C.A. and Fink, G.R. (1985). Ty Elements Transpose through an RNA Intermediate. Cell 40, 491-500.

Britten, R.J. and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science 165, 349-357.

Britten, R.J. and Kohne, D.E. (1968). Repeated sequences in DNA. Science 161, 529-540.

Burton, F.H., Loeb, D.D., Voliva, C.F., Martin, S.L., Edgell, M.H. and Hutchison, III, C.A. (1986). Conservation Throughout Mammalia and Extensive Protein-Encoding Capacity of the Highly Repeated DNA Long Interspersed Sequence One. J.Mol.Biol. 187, 291-304.

Calos, M.P. and Miller, J.H. (1980). Transposable Elements. Cell 20, 579-595.

Corneo, G., Ginelli, E. and Polli, E. (1970). Repeated sequences in human DNA. J.Mol.Biol. 48, 319-327.

Davidson, E.H. and Britten, R.J. (1973). Organization, Transcription, and Regulation in the Animal Genome. The Quart.Rev.of Biol. 48, 565-613.

Deininger, P.L., Jolly, D.J., Ruben, C.M., Friedman, T. and Schmid, C.W. (1981). Base Seuqence Studies of 300 Nucleotide Renatured Repeated Human DNA Clones. J. Mol. Biol. 151, 17-33.

Della Favera, R., Gelmann, E.P., Gallo R.C. and Wong-Staal, F. (1981). A human onc gene homologous to the transforming gene (v-sis) of simian sarcoma virus. Nature 292, 31-35.

Devilee, P., Slagbloom, p., Cornelisse, C.J. and Pearson, P.L. ((1986). Sequence heterogeneity within the human alphoid repetitive DNA family. Nuc.Acids Res. 14, 2059-2073.

Doering, J.L., Jalachich, R. and Hanlon, D.M. (1982). Identification and genomic organization of human tRNA Lys genes. FEBS Lett. 146, 47-51.

Doering, J.L. and Burket, A.E. (1985). Two new families of human repetitive DNA. J.Cell Biol. 101:73a.

Doering, J.L., Burket, A.E., Hanlon, D.M. and Schlegel, D.S. (1986). New Subfamilies of Human Alphoid Repetitive DNA. J.Cell Biol. 103: 491a.

Donehower, L. and Gillespie, D. (1980). Restriction Site Periodicities in Highly Repetitive DNA of Primates. J.Mol.Biol. 134, 805-834.

Fittler, F. (1977). Analysis of the alpha-Satellite DNA from African green monkey Cells by Restriction Nucleases. Eur.J.Biochem. 74, 343-352.

Fitzgerald-Hayes, M., Clarke, L. and Carbon, J. (1982). Nucleotide Sequence Comparisons and Functional Ayalysis of Yeast Centromere DNAs. Cell 29, 235-244.

Flavell, A.J. and Ish-Horowicz, D. (1981). Extrachromosomal circular copies of the eukaryotic transposable element copia in culterured Drosophilia cells. Nature 292, 591-595.

Fritsch, E.F., Lawn, R.M. and Maniatis, T. (1980). Molecular Cloning and Characterization of the Human Beta-Like Globin Gene Cluster. Cell 19, 959-972.

Fujita, A., Hattori, M., Takenaka, O. and Sakaki, Y. (1987). The L1 family (KnpI family) sequence near the 3′ end of human beta-globin gene may have been derived from an active L1 sequence. Nuc.Acids Res. 15, 4007-4020.

Furlong, N.B., Marien, K., Flook, B. and White, J. (1986). Characteristics of Site Variation Among Clones of the 340-Base Pair, Tandemly Repeated EcoRI Family of Human DNA. Biochem.Gent. 24, 71-78.

Gillespie, D., Adams, J.W., Costanzi, C. and Caranfa, M.J. (1982). New orientations of ancestral, "long interspersed repeated sequences" (LINES) in human DNA. Gene 20, 409-414.

Gosden, J.R., Mitchells, A.R., Buckland, R.A., Clayton, R.P. and Evans, H.J. (1975). The location of four human satellite DNAs on human chromosomes. Exp.Cell Res. 92, 148-158.

Gray, K.M., White, J.W., Costanzi, C., Gillespie, D., Schroeder, W.T., Calabretta, B., and Saunders, G.F. (1985). Recent amplification of an alpha DNA in humans. Nuc.Acids Res. 13, 521-535.

Grimaldi, G. and Singer, M.F. (1983). Members of the KpnI family of long interspered repeated sequences join and interrrupt alpha-satellite in the monkey genome. Nuc.Acids Res. 11, 321-338.

Grosveld, F.G., Dahl, H.M., de Boer, E. and Flavell, R.A. (1981). Isolation of Beta-globin-related genes from a human cosmid library. Gene 13, 227-237.

Hattori, M., Hidaka, S. and Sakaki, Y. (1985). Sequence analysis of a KpnI family member near the 3′ end of human beta-globin gene. Nuc.Acids Res. 13, 7813-7827.

Hattori, M., Kuhara, S., Takenaka, O. and Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature 321, 625-628.

Hess, J., Perez-Stable, C., Wu, G.J., Weir, B., Tinoco Jr., I and Shen, C.-K.J. (1985). End-to-end Transcription of an Alu Family Repeat. J.Mol.Biol. 184, 7-21.

Hwu, H.R., Roberts, J.W., Davidson, E.H. and Britten, R.J. (1986). Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proc.Natl.Acad.Sci. 83, 3875-3879.

Jabs, E.W., Meyers, D.A. and Bias, W.B. (1986). Linkage Studies of Polymorphic, Repeated DNA Sequences in Centromeric Regions of Human Chromosomes. Am.J.Hum.Genet. 38, 297-308.

Jagadeeswaran, P., Forget, R. G. and Weissman, S. J. (1981). Short Interspersed Repetitive DNA Elements in Eucaryotes: Transposable DNA Elements Generated by Reverse Transcription of RNA Pol III Transcripts? Cell 26, 141-142.

Jones, R.S. and Potter, S.S. (1985). Characterization of cloned human alphoid satellite with an unusual monomeric construction: evidence for enrichment in HeLa small polydisperse circular DNA. Nuc.Acids Res. 13, 1027-1041.

Jorgenson, A.L., Bostock, C.J. and Bak, A.L. (1986). Chromosome-specific subfamilies with Human Alphoid Repetitive DNA. J.Mol.Biol. 187, 185-196.

Jorgensen, A.L., Bostock, C.J. and Bak, A.L. (1987). Homologous subfamilies of human alphoid repetitive DNA on Different nucleolus organizing chromosomes. Proc.Natl.Acad.Sci. 84, 1075-1079.

Kiyama, R., Hideki, M. and Oishi, M. (1986). A repetitive DNA family (Sau3A family) in human chromosomes: Extrachromosomal DNA and DNA polymorphism. Proc.Natl. Acad.Sci. 83, 4665-4669.

Kurnit, D.M. and Maio, J.J. (1973). Subnuclear redistribution of DNA species in confluent and growing mammalian cells. Chromosoma 42, 23-36.

Lai, E.C., Woo, S.L.C., Dugaiczyk, A. and O'Malley, B.W. (1979). The Ovalbumin Gene: Alleles Created by the Mutations in the Intervening Sequence of the Natural Gene. Cell 16, 210-211.

Lee, T.N.H. and Singer, M.F. (1982). Structural Organization of alpha-satellite DNA in a Single Monkey Chromosome. J.Mol.Biol. 161, 323-342.

Lee, T.N.H. and Singer, M.F. (1986). Analysis of LINE-1 family sequences on a single monkey chromosome. Nuc.Acids Res. 14, 3859-3870.

Lerman, M.I., Thayer, R.E. and Singer, M.R. (1983). Kpn I family of long interspersed repeated DNA sequences in primates: Polymorphism of family members and evidence for transcription. Proc.Natl.Acad.Sci. 80, 3966-3970.

Lewin, B. (1980). Gene Expression 2, Second Edition Eucaryotic Chromosomes. John Wiley and Sons, New York, N. Y.

Maio, J.J. (1971). DNA Strand Reassociation and Polyribonucleotide Binding in the Afican green monkey Cercopithecus aethiops. J.Mol.Biol. 56, 579-595.

Maio, J.J., Brown, F.L. and Musich, P.R. (1977). Subunit structure of chromatin and the organization of eukaryotic highly repetitive DNA. J.Mol.Biol. 117, 637-655.

Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982). Molecular Cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N. Y.

Manuelidis, L. (1978a). Complex and Simple Sequences in Human Repeated DNAs. Chromosoma 66, 1-21.

Manuelidis, L. (1978b). Chromosomal Localization of Complex and Simple Repeated Human DNAs. Chromosoma 66, 23-32.

Manuelidis, L. (1982). Nucleotide sequence definition of a major repeated human DNA, the Hind III 1.9 kb family. Nuc.Acid Res. 10, 3211-3219.

Manuelidis, L. and Ward, D. C. (1984). Chromosomal and nuclear distribution of the Hind III 1.9 kb human DNA repeat segment. Chromosoma 91, 28-38.

Manuelidis, L. and Wu, J.C. (1978). Homology between human and simian repeated DNA. Nature 276, 92-94.

Maxam, A.M. and Gilbert, W. (1980). Sequencing end-labelled DNA with base-specific chemical cleavages. Methods Enzymol. 65, 449-560.

McClintock, B. (1984). The Significance of Responses of the Genome to Challenge. Science 226, 792-801.

McCutchan, T., Hsu, H., Thayer, R.E. and Singer, M.F. (1982). Organization of African Green Monkey DNA at Junctions between alpha-Satellite and other DNA Sequences. J.Mol.Biol. 157, 195-211.

Meneveri, R., Agresti, A., Della Valle, G., Talarico, D., Siccardi, A.G. and Ginelli, E. (1985). Identification of a Human Clustered G + C-rich DNA FAmily of Repeats (Sau3A Family). J.Mol.Biol. 186, 483-489.

Mitchell, A.R., Gosden, J.R. and Miller D.A. (1985). A cloned sequence, p82H, of the alphoid repeated DNA family found at the centromeres of all human chromosomes. Chromosoma 92, 369-377.

Miyake, T., Migita, K. and Sakaki, Y. (1983). Some KpnI family members are associated with the Alu family in the human genome. Nuc.Acids Res. 11, 6837-6846.

Moore, E.E., Jones, C., Kao, F-T. and Oates, D.C. (1977). Synteny between Glycinamide Ribonucleotide Synthetase and Superoxide Dismutase (Soluble). Am.J.Hum.Genet. 29, 389-396.

Musich, P.R. and Dykes, R.J. (1986). A long interspered (LINE) DNA exhibiting polymorphic patterns in human genomes. Proc.Acad.Natl.Sci. 83, 4854-4858.

Orgel, L.E. and Crick, F.H.C. (1980). Selfish DNA: the ultimate parasite. Nature 284, 604-607.

Pardue, M.L. and Gall, J.G. (1970). Chromosomal localisation of complex and simple repeated human DNAs. Chromosoma, 83, 103-125.

Paulson, K.E., Deka, N., Schmid., Misra, E., Schindler, C.W., Rush, M.G., Kadyk, L. and Leinwand, L. (1985). A transposon-like element in human DNA. Nature 316, 359-361.

Potter, S.S. (1984). Rearranged sequences of a human Kpn I element. Proc.Natl.Acad.Sci. 81, 1012-1016.

Prosser, J., Frommer, M., Paul, C. and Vincent, P.C. (1986). Sequence Relationships of Three Human Satellite DNAs. J.Mol.Biol. 187, 145-155.

Reed, K.C. and Mann, D.A. (1985). Rapid transfer of DNA from agarose gels to nylon membranes. Nuc.Acids Res. 13, 7207-7221.

Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977). Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J.Mol.Biol. 113, 237-251.

Rosenberg, H., Singer, M. and Rosenberg, M. (1978). Highly Reiterated Sequence of SIMIANSIMIANSIMIANSIMIANSIMIAN Science 200, 394-399.

Schmid, C.W. and Jelinek, W.R. (1982). The Alu Family of Dispersed Repetitive Sequences. Science 216, 1065-1070.

Schwartz, D., Rizard, R. and Gilbert, W. (1983). Nucleotide Sequence of Rous Sarcoma Virus. Cell 32, 853-869.

Seiki, M., Hattori, S., Hirayama, U. and Yoshida, M. (1983). Human adult T-cell leukemia virus: Complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc.Natl. Acad. Sci. 80, 3618-3622.

Shafit-Zagardo, B., Brown F.L., Maio, J.J. and Adams, J.W. (1982a). Kpn I families of long, interspersed repetitive DNAs associated with the human Beta-globin gene cluster. Gene 20, 397-407.

Shafit-Zagardo, B., Maio, J.J. and Brown, F.L. (1982b). KpnI families of long, interspersed repetitive DNAs in human and other primate genomes. Nuc.Acids Res. 10, 3175-3193.

Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J. and Maio, J.J. (1983). Transcription of the Kpn I families of long interspered DNAs in human cells. Nature 304, 277-280.

Shimizu, Y., Yoshida, K., Ren, C., Fujinaga, K. Rajagopalan, S. and Chinnadurai, G. (1983). Hinf family: a novel repeated DNA family of the human genome. Nature 302, 587-590.

Shinnick, T.M., Lerner, R.A. and Sutcliffe, J.G. (1981). Nucleotide sequence of Moloney murine leukemia virus. Nature 26, 11-17.

Sims, M.A., Doering, J.L. and Hoyle,H.D. (1983). DNA methylation patterns in the 5S DNAs of Xenopus laevis. Nuc. Acid Res. 11, 277-290.

Singer, D.S., (1979). Arrangement of a Highly Repeated DNA Sequence in the Genome and Chromatin of African Green Monkey. J.Mol.Biol. 254, 5506-5514.

Singer, M.F. (1982). SINEs and LINEs: Highly Repeated Short and Long Interspersed Sequences in Mammalian Genomes. Cell 28, 433-434.

Singer, M.F. and Skowronski, J. (1985). Making sense out of LINEs: long interspersed repeat sequences in mammalian genomes. TIBS 10, 119-122.

Smith, G. P. (1976). Evolution of Repeated DNA Sequences by Unequal Crossover. Science 191, 528-535.

Soares, M.B., Schon, E. and Efstratiadis, A. (1985). Rat LINE1: The Origin and Evolution of a family of Long Interspersed Middle Repetitive DNA Elements. J.Mol.Evol. 22, 117-133.

Spradling, A.D. and Rubin, G.M. (1982). Transposition of Cloned P Elements into Drosophila Germ Line Chromosomes. Science 218, 341-347.

Strauss, F. and Varshavshy, A. (1984). A Protein Binds to a Satellite DNA Repeat at Three Specific Sites That Would Be Brought into Mutual Proximity by DNA Folding in the Nucleosome. Cell 37, 889-901.

Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L. and Leinwand, L. (1984). Non-Alu family interspersed repeats in human DNA and their transcriptional activity. Nuc.Acids Res. 12, 2669-2690.

Temin, H.M. (1980). Origin of Retroviruses from Cellular Moveable Genetic Elements. Cell 21, 599-600.

Tyler-Smith, C. and Brown, W.R.A. (1987). Structure of the Major Block of Alphoid Satellite DNA on the Human Y Chromosome. J.Mol.Biol. 195, 457-470.

Van Arsdell, S.W., Denison, R.A., Bernstein L.B. and Weiner, A.M. (1981). Direct Repeats Flank Three Small Nuclear RNA Pseudogenes in the Human Genome. Cell 26, 11-17.

Varmus, H.E. (1982). Form and Function of Retroviral Proviruses. Science 216, 812-820.

Waye, J.S. and Willard, H.F. (1985). Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X Chromosome. Nuc.Acids Res. 13, 2731-2743.

Willard, H.F. (1985). Chromosome-Specific Organization of Human Alpha Satellite DNA. Am.J.Hum.Genet. 37, 524-532.

Willard, H.F., Waye, J.S., Skolnick, M.H., Schwartz, C.E., Powers, V.E. and England, S.B. (1986). Detection of restriction fragment length polymorphisms at the centromeres of human chromosomes by using chromosome-specific alpha satellite DNA probes: Implications for development of centromere-based genetic linkage maps. Proc.Natl.Acad.Sci. 83, 5611-5615.

Wolfe, J., Darling, S.M., Erickson, R.P. Craig, I.W., Buckle, V.J., Rigby, P.W.J., Willard, H. F. and Goodfellow, P.N. (1985). Isolation and Characterization of an Alphoid Centromeric Repeat Family from the Human Y Chromosome. J.Mol.Biol. 182, 477-485.

Wu, J.C., and Manuelidis, L. (1980). Sequence Definition and Organization of a Human Repeated DNA. J.Mol.Biol. 142, 363-386.

APPROVAL SHEET

The thesis submitted by Susan Leah Carnahan has been read and approved by the following committee:

Dr. Jeffrey L. Doering, Director
Associate Professor, Biology, Loyola

Dr. Howard Laten
Associate Professor, Biology, Loyola

Dr. John Smarrelli
Assistant Professor, Biology, Loyola

The final copies have been examined by the director of the thesis and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given final approval by the Committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirement for the degree of Master of Science.

3-20-88
Date

Jeffrey L. Doering
Director's Signature